National Oceanic and Atmospheric Administration (NOAA)

National Environmental Satellite, Data, and Information Service (NESDIS)

Comprehensive Large Array-data Stewardship System (CLASS)

CLASS Architecture Study

February 14, 2003

Review & Approval

Reviewer	Version #	Signature	Date
CMPT	2nd draft		12/01/2002
SET	Final Draft		01/31/2003

Table of Contents

List of Figures

Revision History

Version	Description	Revision By	Revision Date
1st Draft	Initial Study Draft	Vic Church	11/06/02
2nd Draft	Update initial draft with TBDs	Vic Church	11/18/02
Final Draft	Update 2nd Draft with reviewed comments received from CPMT & SET members	Aiman Nachawati	12/30/02
Version 1	Final Edit	Richard Moore	2/14/03

1. INTRODUCTION

The Comprehensive Large Array-data Stewardship System (CLASS) is the National Oceanic and Atmospheric Administration's (NOAA's) premier mechanism for safely and securely archiving large-volume environmental data and data products and for making these data available to researchers, commercial users, and the public. The volumes of data collected by satellite-based instruments and observation systems would overwhelm existing archive and distribution systems without a dramatic expansion in capacity: CLASS is the vehicle for the necessary expansion. At the same time, new observation systems and new derivative data products will be used by an expanding customer base with varied requirements for discovering and retrieving data: CLASS encompasses this expanded accessibility and delivery, as well.

With advances in technology, including faster network access, web-based user interfaces, and sophisticated data analytics, NOAA's National Environmental Satellite, Data, and Information Service (NESDIS) can increase the value of the data collections by providing common-access mechanisms with general and specific data discovery techniques. Information derived from the data can be packaged in new forms. With common-access mechanisms, research tools can integrate multiple data sources such as satellite-based images with buoy-recorded ocean temperatures or fire-fighting campaigns. As the volume of data increases, both the potential and the challenge of data stewardship increase as well.

NOAA/NESDIS currently collects observation data, stores it, distributes it to user communities, and develops data discovery and analysis tools. These collection-storage-distribution efforts are spread throughout the agency, primarily at the NOAA National Data Centers (NNDC's). Developing and planned observation campaigns will deliver vast amounts of additional data, and the new types of data will enhance the value and the applicability of existing data collections. The flood of new data requires extensive data archiving and distribution capabilities, and the opportunities for synergy among data collections demand better coordination between each CLASS data producers and the CLASS management.

This study describes the architecture for CLASS as it is currently being implemented and as it is planned for future consolidated data services. It describes the vision and strategy behind CLASS, identifies the business drivers to be addressed by the evolving system, and depicts the elements of CLASS at the conceptual, logical, and physical levels. This architecture study identifies the organizational aspects of the system (e.g., governance and operations) as well as the functional and data components. It describes the elements that will become part of CLASS, describes the target architecture as a basis for strategic planning, and analyzes transitional activities required for the evolution.

This study refers to design-level documentation including requirements and system descriptions. It does not include detailed requirements or specific hardware and software component descriptions.

1.1 Purpose and Scope

"Information system architecture" provides an integrated, high-level characterization of the components and interactions of a system. This architecture defines the purpose and goals of the system as they are defined, agreed upon by the planners and owners, and governed by the needs of users and operators of the system. It presents technological solutions to information

technology problems. It provides a framework wherein detailed software development and systems integration efforts can achieve near- and long-term goals. An architecture ties the business drivers of the system to the components and functions and operational procedures that comprise the system.

A system architecture achieves the following:

a. Provides a coherent abstraction of the operational elements so that high-level interactions and data flows can be identified.

b. Supports the definition of common, reusable capabilities

c. Supports the definition of system-level tests.

d. Provides a planning tool that permits planners to evaluate both the impact of changes to high-level requirements and the availability of new technology to perform existing functions.

e. Addresses issues of performance goal-setting and monitoring.

"Information system architecture" supports information technology investment management by driving the business case for a system and defining metrics for use in assessing system effectiveness. It is used to align a system with the underlying information technology architecture (ITA) of an organization and to contribute to the evolution of such an ITA. It supports budgetary analysis and response to agency and/or Office of Management and Budget (OMB) direction.

This study provides "system architecture" for CLASS, capturing the context and drivers within NOAA and NESDIS and describing the plan for accomplishing those objectives. This study promotes continued integration among the NESDIS systems that CLASS supports and focuses attention on unresolved issues regarding CLASS. Since CLASS is developed in an evolutionary manner over time, based on campaign priorities and available resources, the architectural details are refined and revised accordingly as the system evolves.

1.2 Intended Audience

This study is intended for managers and developers of CLASS components and for managers and developers of interrelated systems. It provides a discussion of CLASS vision and strategy to assist with planning. It describes the services model for interaction with external data systems. It provides a roadmap for integration of various existing information systems. It captures the high-level business drivers and mission requirements that CLASS is expected to meet.

1.3 Background

New satellite observation campaigns are being prepared for launch and operations. The volumes of data to be collected by these campaigns dwarf the data streams managed by existing archive and distribution systems within NESDIS. The size, number, and frequency of datasets to be stored and distributed require significant expansion of capacity for ingesting, moving, storing, processing, and distributing data. New and continuing remote-sensing campaigns include GOES, POES, (including DMSP), NEXRAD, EOS, NPP, NPOESS, and Metop; numerous in situ observation programs also contribute to the information processing challenge. The CLASS concept was developed as a framework to provide integrated data support while accomplishing the needed capacity expansion.

CLASS is built on a combination of existing information storage and access systems. The primary basis for large-volume data storage and management is the Satellite Active Archive (SAA), which provides a comprehensive archive, access, and distribution capability for POES data, and some derived environmental data products. Covering a broader range of derived environmental data products at smaller volumes, the NOAA Virtual Data System (NVDS) supports user access and ordering for data from many research projects and observation efforts. The NOAA server metadata repository and web site support the discovery of information and provide pointers to specific data systems based on standard metadata. The NESDIS network infrastructure will provide support for data transmission. CLASS may support existing data archives through services and a set of common capabilities.

Managed by a multi-center CLASS Project Management Team (CPMT) under the direction of the NESDIS Chief Information Officer (CIO), CLASS is a key project within NESDIS, drawing on support from several of the NNDCs. The project has been underway since November 2001 and fielded its first operational release in July 2002 (Release 0). Architectural and engineering studies were performed to characterize the architecture of the Archive, Access, and Distribution Segment of CLASS and to select hardware and system software for the high-volume hierarchical data storage systems.

1.4 Related Information

a. Comprehensive Large Array-data Stewardship System (CLASS) IT Architecture Description, CSC (for NOAA NESDIS), July 20, 2001

b. Comprehensive Large Array-data Stewardship System (CLASS) Concept of Operations for the Archive, Access and Distribution Component, October 2002

c. Comprehensive Large Array-data Stewardship System (CLASS) System Requirements, October 2002

d. NESDIS Information Technology Architecture (draft), NOAA NESDIS, July 1, 2002

e. NESDIS Information Technical Architecture (ITA) draft (updated March 8, 2002)

f. Federal Enterprise Architecture Framework, Version 1.1, Federal CIO Council, September 1999

g. Enterprise Architecture: A Framework, Zachman, John A., Zachman Institute, www.zifa.com, 2002

h. Comprehensive Large Array-data Stewardship System (CLASS) Baseline Business Case, CSC (for NOAA NESDIS), July 31, 2002

i. The Nation's Environmental Data: Treasures at Risk; Study to Congress on the Status and Challenges for NOAA's Environmental Data Systems, NOAA, August 2001

j. "Searching for NOAA Data", Earth System Monitor, Vol 10:4, NOAA, June 2000

k. NESDIS Business Area Architecture Study for the Comprehensive Large Array-data Stewardship System (CLASS) Version 1.0: CSC (for NOAA NESDIS), May 2, 2002

l. Comprehensive Large Array-data Stewardship System (CLASS) Archive and Distribution System Architecture Assessment Study, CSC (for NOAA NESDIS), June 25, 2002

m. Satellite Active Archive System and Software Description, library.saa.noaa.gov, 2002.

n. NODC Digital Archive Procedures Revision 1.3 (draft), NOAA NODC, May 2002

o. Section 508 Guidelines, 1998 Federal employees with disabilities access to office systems and information guidelines. http://www.section508.gov/

p. National Oceanic and Atmospheric Administration (NOAA) Privacy Guidelines, http://www.noaa.gov/privacy.html

q. J2EE™ technology and its component-based model simplify enterprise development and deployment, http://java.sun.com/j2ee/

r. Simple Object Access Protocol (SOAP) 1.1, http://www.w3.org/TR/SOAP/

s. Universal Description and Discovery Interface (UDDI), http://www-3.ibm.com/services/uddi/

t. Enabling the Extended Enterprise (E3) developed by CSC to support complex system integration and enterprise IT; e3 is a service mark of Computer Sciences Corporation (CSC)

1.5 Document Organization

An architecture study defines the context for a system and describes the constituent elements with their internal and external interfaces. The architecture is composed of various models of the system, such as data models, process models, governance models, and user interaction models. The representations reflect the different perspectives of the audiences for the study. This study follows the typical organization, starting with the rationale for creating the system and proceeding through levels of increasing detail.

Section 1 is this overview.

Section 2 presents the high-level vision and strategy for CLASS and identifies the context (external forces and interests) and operational concepts for CLASS. Section 2 also defines the models used in the architecture and describes organizational considerations of governance and operations.

Section 3 presents the logical architecture, starting with a more detailed analysis of vision and drivers, issues and requirements, and then presenting the architecture models of processes, components and functions, data, applications and systems, and infrastructure.

Section 4 presents the physical architecture and includes the baseline and target architecture models. Section 4 identifies the elements of the system at a level that supports system design and implementation. It includes the application and technology models at a level consistent with the NESDIS ITA and addresses engineering issues such as standards and planning options.

Section 5 provides a framework for the design architecture in identifying the tasks that system designers and implementers need to address.

Section 6 presents a roadmap for evolution from the baseline to the target architecture.

Appendices provide some of the reference and supporting material.

1.6 Glossary

Adapter	Interface software to provide a middleware link to an existing application or datastore
Ancillary	Data Generated Externally
API	Application Program Interface
ARGO	A broad-scale global array of temperature/salinity profiling floats. It is planned as a major component of the ocean observing system. Deployment began in 2000. Conceptually, Argo builds on the existing upper-ocean thermal networks, extending their spatial and temporal coverage, depth range and accuracy, and enhancing them through addition of salinity and velocity measurements.
AVHRR	Advanced Very-High Resolution Radiometer
BRE	Business Rules Engine
Campaign	A program of interrelated observations focused on a specific research topic, usually limited to a specific period of time, often involving multiple sensor systems and scientific disciplines; used particularly in reference to satellite-based remote-sensing programs
CCB	Configuration Control Board
CIO	Chief Information Officer
CLASS	Comprehensive Large Array-data Stewardship System
CM	Configuration Management
COG	CLASS Oversight Group
COMPS	Customer Order Management Processing System
COTS	Commercial off-the-shelf, packaged software components
CPMT	CLASS Project Management Team
CR	Change Request
CSC	Computer Sciences Corporation
CSR	Customer Service Representatives
DA	Data Architecture
DAB	Data Archive Board

data collection	A set or series of related observations and derived products, often from a single instrument or sensor, with a single high-level description and common characteristics; often open-ended; comprised of one or more data streams
dataset	A set of observations or data products packaged for ease of management, storage, retrieval, and application; for example, the set of scan lines from a satellite observation system that comprise a single orbit might be packaged as a dataset. A data stream (for archive purposes) typically consists of a series of datasets.
data stream	A set or series of observations or data products with a consistent, largely uniform format, structure and metadata; for example, Level 0 and Level 1b data from one sensor would constitute two separate data streams in a single data collection
DBMS	Database Management System
DD	Data Discover
DM	Data Manager
DMSP	Defense Meteorological Satellite Program
DP	Data Processing
e3	An architecture framework ("enabling the extended enterprise") developed by CSC to support complex system integration and enterprise IT; e3 is a service-mark of Computer Sciences Corporation (CSC)
EOS	Earth Observing System
external repository	Observation data management system residing on infrastructure external to CLASS, including those operated by NNDC personnel (i.e., within NOAA); CLASS may provide services to such repositories and may include metadata from them in the CLASS portal.
FEAF	Federal Enterprise Architecture Framework
FGDC	Federal Geographic Data Committee
FTP	File Transfer Protocol
GIF	Graphics Interchange Format
GIS	Geospatial Information System: an information management system that supports operations on data based on the physical location to which that data refers (e.g., latitude, longitude, elevation)
GOES	Geostationary Operational Environmental Satellite
GOTS	Government Off-The-Shelf
GUI	Graphical User Interface
HDF	The National Center for Supercomputing Applications (NCSA) developed Hierarchical Data Format, a standard for complex data

	records and datasets adopted by NASA for satellite data products. HDF supports mixed-format data (e.g., text, binary scan data, GIF images, video) to be treated as an aggregate rather than as distinct elements. Versions HDF-4 and HDF-5 are of current interest.
HDS	Hierarchical Data System
HPSS	High-Performance Storage System
HSM	Hierarchical Storage Management
HTML	HyperText Markup Language
HTTP	HyperText Transfer Protocol
ICD	Interface Control Document
IDTV	Internet Digital Television
IOOS	Integrated Ocean Observing System
IS	Information System
IT	Information Technology
ITA	Information Technology Architecture
J2EE	Java version 2 Enterprise Edition, a framework for integrating server applications using Java interface specifications
JPEG	Joint Photographic Experts Group, a standardized image compression mechanism
LDAP	Lightweight Directory Access Protocol, an interface definition for application interfaces to a user directory; based on the X.500 directory protocol with somewhat reduced functionality
metadata	Information describing other information; metadata comes in several forms and flavors including content metadata (describes the substance of an information set), administrative metadata (describing the form and history of an information set), and access metadata (providing access information for the information set); it can be embedded with the content that it describes or maintained independent of the content.
MDDB	Metadata Database
METOP	Meteorological Ops/EUMETSAT Meteorological Observation Satellite System
Middleware	Messaging system with standard protocols to connect different applications, often using message queues to support asynchronous interactions
MIE Scattering	Radar backscattering by targets having dimensions somewhat greater than 1/10 the wavelength of the radar but less than several radar wavelengths. Any scattering produced by spherical particles without special regard to comparative size of radiation wavelength and particle diameter. Contrasted with Rayleigh scattering.

MIME	Multi-Purpose Internet Mail Extensions
MODIS	Moderate Resolution Imaging Spectroradiometer
MPEG	Moving Picture Experts Group, a standardized coded representation of digital audio-visual information in a digital compressed format compression mechanism.
NCSA	National Center for Supercomputing Applications
NARA	National Archive And Records Administration
NAS	Network Attached Storage
NASA	National Aeronautics and Space Administration
NCDC	National Climatic Data Center
NCSA	National Center for Supercomputing Applications
NESDIS	National Environmental Satellite, Data, and Information Service
Net CDF	Common data format combining raw data and metadata in a single dataset using a predefined structure; adopted by NOAA for data interoperability
NEXRAD	Next Generation Weather Radar
NGDC	National Geophysical Data Center
NIST	National Institute of Standard and Technology
NNDC	NOAA National Data Centers
NOAA	National Oceanic and Atmospheric Administration (of the Department of Commerce)
NOAA Server	NOAA environmental data access metadata repository and portal
NODC	National Oceanographic Data Center
NOMAD	Navy Oceanographic Meteorological Automatic Device
NOS	National Ocean Service
NPOESS	National Polar-Orbiting Operational Environmental Satellite System
NPP	NPOESS Preparatory Program
NVDS	NOAA Virtual Data System
OLS	Operational Line Scanner
OMB	Office of Management and Budget
OMS	Order Management System
OSDPD	Office of Satellite Data Processing and Distribution
Observation Program	A program of observations, often open-ended or episodic, typically involving a single sensor or sensor net, intended to provide environmental data without regard to a specific research

	focus; for example, Great Lakes bathymetry; ocean temperature or salinity measurements
PDA	Personal Digital Assistant
PMEL	Pacific Marine Environment Laboratory
POES	Polar-Orbiting Operational Satellite System
QA	Quality Assurance
RDBMS	Relational Database Management System
SAA	Satellite Active Archive
SAM-FS	Storage and Archive Manager File System
SET	System Engineering Team
SMTP	Simple Mail Transfer Protocol
SNA	Storage Network Architecture
SNMP	Simple Network Management Protocol
SOAP	Simple Object Access Protocol, a “web services” application interface
SST	Sea Surface Temperature
UDDI	Universal Description and Discovery Interface, a directory system for web service publication and location
UI	User Interface
URL	Universal Registry Locator
WAP	Wireless Application protocol
XHTML	Extensible HyperText Markup Language
XML	Extensible Markup Language

2. CLASS VISION, CONTEXT, AND CONCEPT

2.1 Vision

The starting point for the conceptual view is the vision for CLASS: what is CLASS intended to accomplish and over what time period will it meet its goals?

The vision for CLASS is documented in several publications, some of which are included as appendices to this study. These include the vision statement from NESDIS CIO (April 24, 2002) and the NESDIS ITA draft (updated March 8, 2002):

a. CLASS is the premier data storage system for high-volume, large-array environmental data, which is chiefly collected by or derived from satellite-based observations.

b. CLASS focuses on the environmental observation data and related data products collected through NOAA observation campaigns and research activities but not directly concerned with the science characteristics of the data.

c. CLASS provides secure, disaster-proof, efficient, and accessible storage to meet the needs of environmental researchers and other users.

d. CLASS supports smaller volume collections of environmental data, including non-satellite remote sensing and in situ observation data products.

e. CLASS is the entry point for access to all NOAA environmental data through a single portal, providing data discovery services for users with different needs and capabilities including research, educational, commercial, and general public.

f. CLASS facilitates metadata maintenance describing the nation's environmental data assets and supports access to and requests for such data.

g. CLASS fosters the use of advanced data discovery techniques such as visualization and geospatial information systems developed for specific data collections.

h. CLASS provides access to its archives and (with appropriate mechanisms) to other repositories described in the metadata repository.

i. CLASS provides a common set of ordering and customer interface services to other NESDIS data systems.

j. CLASS provides the distribution system for its archives and facilitates distribution of datasets and products from external archives.

k. CLASS is NOT intended to provide campaign- or project-specific data processing services (e.g., creation of level 1b data from level 0 data). It provides limited processing for requested data products (e.g., generating geographically bounded subsets of orbit-based datasets). It provides limited data selection and retrieval capabilities to its archived data. It does NOT replace or incorporate content-specific, research-based data analysis and discovery information systems, but CLASS might provide access to those systems.

l. CLASS performs automated quality assurance (QA) processes on ingested data to ensure readability, detect transmission errors, ensure compliance with dataset format agreements, and assess record quality indicators. It does not perform QA on derivative science data or assessments of utility for research. The processing will be defined specifically for each data stream.

m. CLASS operates and evolves within the NESDIS IT infrastructure and responds to direction (e.g., legislative mandates) from oversight and management groups.

n. CLASS is designed and implemented to incorporate advances in technology that improve the performance of its primary mission.

Appendix A includes base documents from which this vision statement was derived.

2.2 Context View

2.2.1 Planner Perspective

Figure 1 shows the contextual view that includes the *planner oversight* perspective on CLASS.

Figure 1. Business Architecture: Contextual View

The figure characterizes CLASS as a system providing services in response to requirements levied by:

a. Suppliers of environmental data who require data to be stored securely and made available to customers effectively

b. Users/Customers (e.g., customer-service representatives) and customers who want to access data

c. Owners (NESDIS) who define the vision and business drivers for CLASS and need assurance that the provider-customer interchange is effective, secure, and efficient,

d. Planners: operators, developers, and managers of CLASS who must keep it functional while evolving it to address new data requirements.

2.2.2 Interface Context

The previously defined groups (suppliers, customers, owners, and planners) provide the *context* in which CLASS functions. Figure 2 identifies the external entities and stakeholders who contribute to the requirements and planning for CLASS.

Figure 2. CLASS Data and Management Information Context Map

The context map for CLASS includes organizations and individuals who interact operationally with the system (e.g., data suppliers and customers) and those whose roles involve oversight and development. These stakeholders are the primary source of goals, assumptions, constraints, and functional requirements for CLASS.

2.2.2.1 Stakeholders

a. **Data Providers**. Researchers create or capture data products intended for archive and dissemination and deliver them to CLASS, along with appropriate content metadata. NOAA's NNDCs, NASA researchers, or specific observation programs and campaigns generate (or pass through) such information via direct interfaces to CLASS. Information from outside this context (e.g., from foreign partners) is typically processed through a program in one of the NNDCs.

Data providers also supply application elements such as functions and tools to support specific format-based processing such as product quality verification and creation of browse imagery. These may include scripts, tables, or application code to support the CLASS end of the interface. Those functions and tools will be hosted in CLASS as a service, and any maintenance or upgrade to these functions and tools are the responsibilities of the data providers.

A primary source of specific data volume and access requirements levied on CLASS, data providers supply input to format and protocol selections. However, they do not mandate which requirements ultimately get incorporated into CLASS.

b. **Oversight**. Overseen by the CLASS Oversight Group (COG), CLASS responds to the direction of the Data Archive Board (DAB). Under the auspices of the NESDIS Office of the CIO, the CLASS Project Management Team (CPMT) provides direction for the CLASS project. The CPMT includes representation from each of the NNDCs involved in development and operations of CLASS. Section 2.6, Organizational Considerations—Governance, discusses the roles of these oversight groups.

c. **Research Customers**. The primary customers for CLASS-archived data are researchers in various organizations. Many of these are researchers from NOAA projects or non-mission-critical operational data users. Others are academic researchers or research communities. In some cases, the research customers are also data providers, creating data products from archived data and delivering those products for permanent storage and/or dissemination.

   Research customers may originate requirements for particular distribution mechanisms and formats and for data discovery and display capabilities. Because of the need for timely and repeated data distributions to support temporal sequences, research customers may request subscription product orders and automated interfaces to CLASS archives.

d. **Citizen Customers**. General citizen access to environmental data (including that archived in CLASS) is less comprehensive and predictable than research customer requirements. The nation's environmental data collections support academic users (typically teacher-directed discovery and research) and general public access to information, studies, and images. The primary interaction channel is Internet and web-based. Requirements for supporting novice and naïve users are primarily directed at these customers.

e. **Commercial Customers**. Businesses make use of NOAA environmental data in much the same way as research customers, although this category is less likely to generate data products directly for CLASS storage. Many commercial customers use CLASS-provided environmental data for the creation of value-added products. Commercial customers may have requirements for special services (e.g., certification of data products for legal purposes) and may require large volumes of data.

f. **Implementers**. Implementation of CLASS is a slightly ambiguous category, half in and half out of the CLASS context boundary. The CLASS team implements the core system internally, while data providers will also be allowed to provide application elements such as functions and tools to support specific, format-based processing. (CLASS *operations*, in contrast are considered to be completely within CLASS.) Implementers for different elements of CLASS can be part of other categories (e.g., data providers and research customers) and act as reviewers and requirements-providers as well as developers.

The implementation teams are elements of NESDIS and include personnel from the Office of Satellite Data Processing and Distribution (OSDPD) and the Climatic, Oceanographic, and Geophysical data centers.

g. **Policy and Standards Organizations**. Various policy-making and standard-setting organizations define the constraints on CLASS. These constraints include development and management control exercised by NOAA and OMB, implementation standards defined by the National Institute of Standards and Technology (NIST) and the Federal Geographic Data Committee (FGDC), archive management rules specified by the National Archives and Records Administration (NARA), and operational standards including those defined by NASA. The CLASS project plays an active role in some of these organizations (e.g., FGDC) as well as conforming to guidelines.

h. **Metadata Interfaces**. In addition to the data products archived and distributed through CLASS are data elements for which CLASS provides a portal and discovery services. These products may reside outside of the CLASS system where CLASS, as a portal system, provides the users the necessary access and services to connect and order data. These capabilities involve receipt and management of product metadata with references to actual data stored elsewhere. CLASS interfaces with the FGDC Clearinghouse as well as other metadata providers to simplify access to environmental data.

2.2.2.2 Context Interface Elements

The following table identifies the major information elements with which each stakeholder group is substantially involved. The list is not exhaustive but provides the basic information flow. The "concerns" column refers to aspects of CLASS expectations with which each stakeholder group is concerned.

Stakeholder	To CLASS	From CLASS	Concerns
Data providers	a. Observation data products (initial and derived) b. Product metadata c. Ancillary data documentation software d. Interface applications	Reports	Data volumes Ingest interfaces Archive security ICDs
Oversight	Direction	Studying	All
Research customers	a. Discovery interactions b. Orders for data c. Subscriptions d. Automated retrieval transactions	Discovery information Data products	Discovery tools Automated interfaces Consistency of format and channel Formats and protocols

Stakeholder	To CLASS	From CLASS	Concerns
Citizen customers	Discovery interactions Orders	Discovery information Data products	Ease of use (GUI) Timeliness Accessibility
Commercial customers	a. Discovery interactions b. Orders for data c. Subscriptions d. Automated retrieval transactions	Discovery information Data products	Discovery tools Automated interfaces Special services ICDs
Implementers	Application solutions		All
Policy and standards organizations	a. Standards b. Guidelines Processes	Input to standards and guidelines	Standards and guidelines
Metadata interfaces	a. Content metadata b. Access metadata c. Metadata maintenance interactions	CLASS metadata	ICDs

2.3 Business Drivers

Business drivers are the goals to be accomplished, the strategic direction that drives specific solutions from competing possibilities, and the constraints that restrict the range of solutions and the scope of the project. For CLASS, these can be stated as follows.

2.3.1 Goals

CLASS is to accomplish the following:

a. Provide secure storage for massive amounts of data (anticipated to approach 15 PB [Petabytes] by 2015, growing from slightly more than 30 TB [terabytes] currently stored at the SAA

b. Provide reliable 24/7 data receipt, evaluation, and acknowledgment for continuing data streams (as specified in negotiated interface control documents with suppliers)

c. Support the GOES, POES (including DMSP), NEXRAD, EOS, NPP, NPOESS, and Metop campaigns

d. Support archiving of and access to in situ observation data

e. Support archiving of and access to derived products

f. Distribute secure data (including data denial aspects, as necessary) and timely manner on request via appropriate mechanisms (Internet, e-commerce and physical media)

g. Facilitate discovery of and access to NOAA environmental data

h. Support an increasingly diverse set of customers as they explore available data and request selected data products

2.3.2 Design Drivers

The CPMT identified and prioritized the following set of system characteristics that the CLASS implementation must reflect. Section 4.3 discusses these characteristics. They are listed here in priority order as determined by the CPMT.

#	Characteristic	Description
1	Reliability	The architecture provides dependable services for users. A combination of Correctness (Ability to perform their exact tasks, as defined by their specification) and Robustness (Ability to react appropriately to abnormal conditions).
2	Availability	The architecture supports operational capabilities that are consistent with the Service Level Agreements established for this application. If the supporting infrastructure goes down, it should be able to recover and restart the applications within acceptable time frames.
3	Usability	The architecture supports ease of use so that a user can learn to operate, prepare inputs for, and interpret outputs of a system or component.
4	Extensibility/Extendibility	The architecture supports adaptation to changes in specification. Otherwise stated: The architecture is easily modified to support new business functions.
5	Interoperability	The architecture supports the ability of two or more systems to exchange information and to use the information that has been exchanged.
6	Scalability	The architecture can grow to provide acceptable levels of performance for multiple users and multiple resources at specified transaction volumes. The scalability can be provided at the hardware, system-level software, and application-level software.
7	Modularity	The architecture supports discrete components such that a change to one component has minimal impact on the other components.
8	Modifiability	The architecture facilitates the incorporation of changes, once the nature of the desired change has been determined. These criteria relate to existing system hardware and software; this is different from flexibility that pertains to new technologies and extensibility which pertains to changes in specifications or business functions.
9	Reusability	The architecture supports components that can serve for the construction of many different applications.
10	Integrity	The architecture provides protection against unauthorized access and modification.
11	Manageability	The architecture supports tools and/or custom software that will simplify the systems management functions for components of the architecture. Stated otherwise, the architecture supports the ease of operation of a deployed system. It involves the administrative use of the system.

#	Characteristic	Description
12	Testability	The architecture facilitates development of test criteria and performance tests to determine whether those criteria have been met. Testing includes operational correctness, robustness, and efficiency. Not only is testability a measurement for software, it can also apply to the testing scheme.
13	Reparability	The architecture supports the ability to facilitate the repair of defects.
14	Understandability	The architecture defines the components, their relationships to other components, and how components interact. Understandability also encompasses the degree to which the purpose of the system or component is clear to an evaluator.
15	Flexibility	The ability of the architecture to accommodate and incorporate new technologies as they become available. Another definition is that the architecture supports the ability to easily modify the system or components for use in applications or environments other than those for which it was specifically designed.
16	Efficiency	The architecture supports a system or component that performs its designated functions with minimum consumption of resources.
17	Recoverability	The architecture provides error detection and recovery functions for all of its services.
18	Security	The architecture provides the ability to manage, protect, and distribute sensitive information. Security also encompasses the ability to manage user login and system access.
19	Portability	The architecture supports the transfer of system or components from one hardware or software environment to another.
20	Deployability	The architecture supports the ease of system deployment. Deployment includes acquisition of hardware and software package, site preparation, installation, configuration, integration, testing, training of on-site staff. Deployment applies to the targeted hardware or software environment.
21	Survivability	The architecture provides that essential functions are still available even though some part of the system is down.

2.3.3 Strategic Direction

The strategic context for CLASS development includes the following:

a. Using networks and the Internet for data transfers and distribution (as much as practicable)

b. Using the SAA as the baseline for the Data Archive, Access, and Distribution functions of CLASS

c. Using NVDS functionality to support the order management function

d. Using NOAA server as the model for the access portal to NOAA environmental data including CLASS-managed archives (e.g., incorporate and expand on the functionality provided by NOAA Server)

e. Providing geographically separated redundant data storage capabilities to effect disaster-proof archiving

f. Focusing on new observation (primarily satellite) campaigns and adding existing repositories in accordance with budgets and priorities

g. Using internet-based technology for interactive user interfaces

2.3.4 Constraints

Development of CLASS must conform to the following constraints.

a. Alignment with the NESDIS ITA

b. Coordination with external data programs (e.g., EOS)

c. Integration with existing archives and data systems

d. Compliance with NARA guidelines for permanent records management

e. Compliance with legislative and OMB direction

f. Provision of appropriate security

g. Compliance with the FGDC standard

h. Compliance with section 508 of the rehabilitation and NOAA privacy guidelines

2.4 Concept View

The context view of a system architecture characterizes the system from the outside; the *concept* view describes the system, without excessive detail, from the inside. In Zachman and Federal Enterprise Architecture Framework (FEAF) sense, it is the "owner's point of view" of a system addressing the major building blocks that respond to external drivers without presenting the internal workings.

2.4.1 CLASS Operations Concepts

The core data flows of stewardship involve moving data products into archival storage and making those products available for identification and retrieval. These core data flows occur in all archive systems, and CLASS exists within an environment that already addresses the issues with which the new system must contend. CLASS provides a powerful response to the massive data quantities that new satellite campaigns will collect and an approach to streamlining and integrating existing archive activities. Figure 3 shows the conceptual data functions and flows and places CLASS within the overall data process.

Figure 3. CLASS Encompasses the Core of Observation Data Management

2.4.2 External Elements

External to the CLASS domain, **suppliers** generate the data products to be archived and made available. (The observation data collection process is not shown.) Elements to be archived may include raw data, but in general consist of level 0 and level 1b data. Other derivative products (e.g., gridded parameters and images created from mosaics of orbit files) may also be archived. To support retrieval and subsequent use of the data products, descriptive information is also provided to characterize the data itself and its processing history. (Although this is "metadata" in a general sense, the term has a quite-specific meaning in the NESDIS context. The term ancillary data is used as a generic term).

Research teams may create archive systems for specific instruments or projects (e.g., models and simulations) and encapsulate the entire data flow in a single data system. Typically these **external archive systems** incorporate substantial data processing (whether reducing raw data or creating derivative products) and include data discovery mechanisms that are highly content-specific. These archive systems may deliver data products to users through a common mechanism (e.g., CLASS) or directly via network and web interfaces.

On the other side of the data flow, **customers** interact with the data system to discover data, request files, and receive data products. Customers may interact solely with CLASS, via CLASS with research-developed archive systems, or directly by web interface with the content specific

web sites. The mechanisms implementing these channels are discussed in the logical and physical architecture descriptions.

Customers may be research teams that create new products from archived observations. These products may be added to the archive with CLASS-management approval, so that customers can also be suppliers to the archive.

2.4.3 CLASS Operational Elements

CLASS serves as the core of the data archive and distribution process for observation data products. It performs four interrelated but distinct functions: archiving and managing data, supporting data discovery, enabling customers to request data products, and filling the orders placed. In doing so, it uses several distinct data stores (one of which is the archive itself).

2.4.3.1 Archiving and Managing Data

Ingest Function. The process of receiving, evaluating, storing, and indexing data is captured in the ingest function. Data products and associated ancillary data are acquired, either through passive receipt or through an active notification and retrieval process. (There is an associated governance process for each data stream to define what data are to be acquired, the formats that are employed, the quality checks to be performed to insure correct receipt, and workflow requirements such as "hold seven days after acknowledging receipt before discarding.") The products are processed in some form to verify data transmission, extract metadata from datasets, link ancillary data with products, adjudicate occurrences such as duplicate (possibly replacement) data products, and create browse imagery or other data discovery artifacts. Products are stored in appropriate locations (e.g., tape or optical storage, online cache, redundant archives).

Distributed Hierarchical Storage Management (HSM). Satellite-generated environmental data is notably voluminous. No current technology allows all of the data and all the derivative data products to be stored online and instantly available. The conceptual solution is to use a hierarchical data storage (HDS) system providing fast access to data most likely to be requested while supporting delayed access to other (typically older) data. Several technological approaches include online, "near-line," and off-line storage in a single virtual storage system. CLASS also provides for distributed archival storage in geographically separate locations. Details of the HDS capability, such as mechanisms and policies for positioning data products, are detailed in the logical and physical architecture discussions.

Archive. The archive consists of two or more storage systems located in different data centers that hold information for archival purposes. The current technology is magnetic tape managed in robotic tape libraries, configured to comply with NARA guidelines for permanent storage of national data records. The archived data is managed through the **distributed HSM** function, but is distinguished because of the specific requirements pertaining to permanent storage systems.

Data Products Management Data Store. CLASS must manage a large and diverse population of data products; several types of information must be collected. The information is managed in a catalog and an inventory system. The **catalog** of data products describes the datasets in terms useful for search and retrieval. The **inventory** identifies where copies of the datasets are located. In cases of physical media copies (e.g., CD-ROMs of commonly requested data suites), the inventory would identify physical media numbers as well as its location. (While CLASS focuses

primarily on electronic storage and delivery, some customer circumstances require physical media to be accommodated.)

The data products management data store is the conceptual repository for all of the information about the archived data products, including **metadata** and other data-discovery related information. The physical implementation is typically a distributed data system, but the data products management data store is conceptually one element.

Publish. The publish function accomplishes the activity of making the availability of data known to processes and users. It involves creating catalog and maintaining data entries for stored products, inventory, and directory information, as well as an interface with standing order systems that can be triggered by the newly processed data. *Catalog information* describes the content of received data; it is drawn from content metadata and supports searching by attributes of the data such as geospatial location, type of measurement, and/or time of measurement. *Inventory information* describes the locations of the datasets or other structures (e.g., relational data tables) in the system. (Note that a dataset may be found in two archives, online cache and CD-ROM discs, all at the same time.) *Directory information* lists datasets in a hierarchical display, comparable to the directory of a file system.

2.4.3.2 Discovering Information

Data Discovery Function. There are three data discovery mechanisms tied to the way that data are managed within the archive. The mechanisms work in concert to provide effective data discovery to customers with a broad range of interests and expertise. The **directory browse** approach uses a hierarchical data structure that supports browsing and automated discovery tools. In cases where a continuing data stream (e.g., Advanced Very High Resolution Radiometer (AVHRR) orbit-based datasets) is represented by a growing list, new data can easily be identified. The catalog and inventory provide dataset-based records. The **metadata search** approach uses descriptive information such as geospatial indicators to identify data products meeting search criteria. Lists of data product classes (e.g., AVHRR data) and/or data products (e.g., a particular composite image) can be generated from the metadata system. **Advanced search** mechanisms are tied to specific content and may be provided by links to supporting repositories (including external data sites).

Supporting these discovery mechanisms are **discovery processing** capabilities that operate on metadata and browse imagery, permitting users to refine their selections through techniques such as low-resolution previews or application of filters to data products. Mechanisms that are completely content-specific will be provided by referral to the appropriate content-specific web sites, while generally applicable mechanisms will be candidates for implementation within the CLASS discovery service.

The data discovery function relies on the Data Processing (DP) Management data store for its underlying indices and information. The associated interfaces provide mechanisms to link to specialized sites and to select and order data products.

2.4.3.3 Requesting Information

Order Function. Customers can place orders for data products in several ways, but the processing requirements are similar in each case. Customers need to provide some registration information (such as an email address for notification of delivery). The **user accounts**

management function supports registration and authorization. Time- and role-based restrictions apply to some data; so customers may be required to provide identification as well. Some data products require payment. A **payment processing** capability, integrated with the accounts management function, provides an ability to prepay accounts or to establish payment procedures. Account information can be maintained and updated for use in ordering different products. Selection of data products can occur in several ways, including manual (web-based) selection from data discovery lists, subscription orders triggered by newly received products meeting established criteria, automated requests generated by programmatic interactions, orders through customer service, or bulk orders that may act as repeating sub-order generators. **Order management** supports the aggregation of data product selections into orders, pricing as appropriate, periodic evaluation of subscriptions, and staged processing of bulk orders. An **order-tracking** capability provides for responses to user inquiries as well as studying for system management and evaluation purposes.

Order Information Data Store. The ordering mechanism is supported by a data system that maintains records of order creation, processing, and fulfillment. (This data system also supports data discovery and other user interactions, but those interactions are better detailed in the logical architecture rather than the conceptual.)

2.4.3.4 Distributing Data

Distribute Function. Completing the archive-access-distribution model is the distribution of data products to requesting customers. Order items identify data products from inventory records, specify any delivery processing, and direct the packaging and delivery of the requested products. Order fulfillment is coordinated with the Ordering function (i.e. through the order info data store) to support tracking, cancellation, and payment activities. The **retrieve or pick** function causes the data products to be staged from their archive locations or (in the case of pre-positioned media) to be picked from storage. Any necessary **delivery processing**, such as subsetting data from orbit datasets, is performed. The retrieved, processed items are **packaged** for shipment (either with a workflow record for electronic transmission or physical packaging and labeling for physical media). Items are **shipped** or staged in a retrieval location (e.g., a File Transfer Protocol (FTP) site) as appropriate, and **notification** of shipment is made. Details of shipping, such as the incremental delivery of bulk orders, are described in the logical and physical architectures.

2.4.4 Interfaces

Suppliers and customers communicate with the archive in several forms. Data and data products are received and shipped through **Internet** or **intranet** services, using protocols such as FTP or secure FTP and other gateway mechanisms. Within the NESDIS network, faster, more flexible protocols may be employed to optimize the use of network bandwidth. Communications about data products may use the same channels or others including interactive web access, direct machine-to-machine gateway connections, email, and interactions with help desk personnel or customer service representatives (CSRs).

The multiplicity of communications channels is required to address the range of customers using the system. Researchers and project teams may require regular deliveries of data as they become available; they are supported with automated ordering and delivery systems. Casual users such as a seventh-grade science class may require much more interaction; they are supported with web-

based interfaces. Retrospective research or data reprocessing efforts may require large amounts of data (bulk orders); mail or phone contact with CSRs provides effective support.

Additional interfaces involve data management interactions with suppliers to define formats and transfer protocols. Some data formats (e.g., Graphics Interchange Format (GIF), Joint Photographic Experts Group (JPEG), Moving Picture Experts Group (MPEG)) require specialized tools for processing. Observation programs may provide tools to read, QA, display, and catalog products employing these formats. Similarly, standard Extensible Markup Language (XML) schemas may be defined for data products and provided to CLASS by suppliers.

2.5 Services Model

One of the ways to view a support system, particularly one that must support an indeterminate number of coordinated systems, is to define the services that the support system provides. CLASS provides ingest-to-distribution processing for some observation programs, but for others (e.g., established programs), CLASS may provide independent elements of that data flow. CLASS features a contiguous set of services for some customers but not for others. Modeling CLASS as a set of service models makes it easier to use a divide and conquer approach, through compartmentalizing the integration aspects.

CLASS is not a monolithic hardware-software system, but an integrated collection of services that facilitate the interaction between data providers and data users. Each of the high-level processes and process-groups is supported by one or more services. These services include archive storage, "shopping cart" user interface and order-management interface, publishing (via standard format metadata) of data products, distribution of data, user profile and account management, credit card processing (through the interface to the NESDIS e-commerce segment), and metadata updates.

This study distinguishes between internal information systems and external information systems, based on where the data primarily resides. CLASS provides all the necessary services for the internal Information Systems (IS); for external systems, CLASS provides a more limited set of published services, as suggested in Figure 4.

Figure 4. Publishing Shared Services for External Information Systems

Details of services (both internal and shared) should be provided in the logical and physical architecture discussions and in the appropriate future system requirements and design documentation.

2.6 Organizational Considerations--Governance

The CLASS architecture represents a set of decisions made that affect the scope and capabilities of the system. Managing and extending CLASS capabilities requires continuing analysis and decision-making concerning all aspects of the system including budget, architecture, scope, and strategic direction. Some of the decision-making concerns the many data repositories to be added to both existing and new observations data products. New technology provides opportunities and requires alignment with other NOAA/NESDIS systems. An expanding customer base requires new interface support and training. Legislative changes regarding data storage and access affect CLASS. The architecture of CLASS includes the organizations and processes involved in overseeing the evolution of the system. Figure 5 shows the primary areas of governance and the roles involved in CLASS governance.

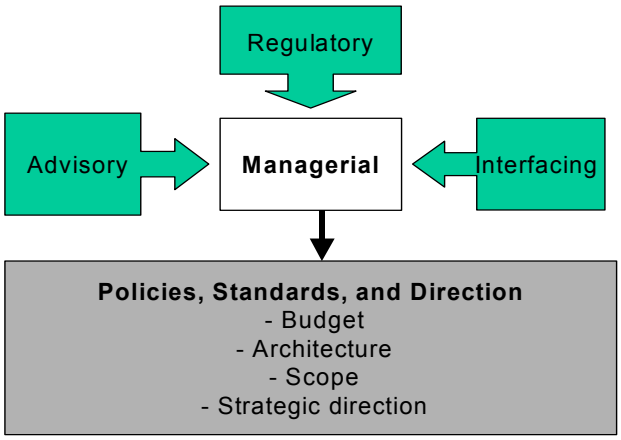


Figure 5. Governance Structure for CLASS Oversight

3. CLASS LOGICAL ARCHITECTURE

This section describes the basic stewardship architecture that is utilized throughout NESDIS (including the SAA, the structural core of CLASS) and the basic data operations model. Section 3.2 discusses the various business drivers that form the basis for requirements. Section 4 describes the current and target architectures.

This section describes a *logical* architecture because it addresses its characteristics, goals, and functional components that support them, but it does not describe the physical implementation of CLASS or the other data systems that are in use within NOAA. The details of actual data management facilities should be addressed in the Design architecture section.

3.1 Extended Vision

The following table expands on the vision provided in Section 2.1 and notes some of the implications for the CLASS architecture.

Item	Element	Implications
1	CLASS provides access to all NOAA data through a single portal.	CLASS provides appropriate data services for data aggregates: a. Owned, managed, and archived in the CLASS repository b. Residing in other archives with distinct UI and distribution capabilities c. Stored in legacy data systems and accessed via the CLASS UI and distribution d. Cost & schedule associated if CLASS provides a single point of entry and consolidated navigation services to support external archives.
2	CLASS archives, manages, and distributes large-array datasets from current and future observation campaigns.	CLASS provides data management services for "owned" data (e.g., physical storage) in the multi-petabyte range. It provides a common services framework so that new campaigns can easily develop data-unique capabilities while leveraging standard interface and distribution capabilities.
3	CLASS provides access and delivery via the Internet.	a. CLASS uses standard Internet protocols including TCP/IP, HTTP, and FTP for access and distribution. b. CLASS includes safeguards commensurate with the risks implied by Internet accessibility. c. Design and development of new capabilities preferentially employ Internet-based technology. d. The CLASS user interface reflects best practices in government and commercial Internet (primarily web) based systems.

Item	Element	Implications
4	CLASS supports data discovery and retrieval that is user-needs-centric rather than campaign-centric.	a. CLASS provides comprehensive directory and search capabilities that address all data collections. b. Data discovery services cross campaigns and data locations. This does not preclude campaign-specific discovery services; it should complement and leverage them.
5	CLASS supports novice and inexperienced users as well as intermediate and advanced users from a growing user base.	a. Within the constraints of Section 508 and privacy guidelines, the CLASS user interface provides a user-selectable form of dialog, providing advanced users with fast, concise access (e.g., to review a standing order) while providing other users with more guidance and assistance. b. The user interface provides multiple means of discovering and ordering data. c. CLASS maintains user-controlled profiles to customize interactions based on preferences and experience.
6	CLASS provides secure data storage with multiple storage locations, backup, and disaster recovery capabilities.	a. CLASS operations include data security planning and procedures to ensure data integrity and recoverability. b. Recovery and disaster response plans are regularly tested and contingency plans updated.
7	CLASS provides redundant, geographically distributed data processing support for ingest, management, discovery, and distribution.	CLASS provides replication of data and data products so that all data are stored in and are accessible from, at a minimum, at least one other location. Network capacity is provisioned to support regular replication among sites. Plans and accommodations are made to provide access (possibly with reduced performance) to all data in the event that one site becomes unavailable.
8	CLASS reduces the need for new observation campaigns to construct unique data processing systems by providing common information services.	CLASS identifies, selects, and/or defines standards, protocols, and services for use by all new campaigns, including: a. Data format standards b. Browse image standards (for data formats and sizes) and guidelines (regarding what images to provide) c. User-authentication and role-definition services (e.g., LDAP support) d. Network-protocol standards including security protocols e. Large-scale data management f. Order management for free- and cost-based orders, and electronic and physical delivery g. Interfaces to financial systems h. Distribution services i. User interface services To maximize the longevity of developed systems, widely accepted and public standards are used where practical.

Item	Element	Implications
9	CLASS supports new campaigns as they become operational and incorporates existing data systems (directly or with managed access) as priorities and resources dictate.	a. CLASS is designed for flexibility to incorporate new directory and search information and new data types as required to add new collections of data. b. It has the capacity to expand data storage, backup, and communications resources to support new collections and data systems. The CLASS project works with observation projects to forecast volumetric and functional requirements and to evaluate options for meeting those requirements. c. For existing data systems, the CLASS project contributes to cost-benefit analyses of migrating some or all data and functionality to CLASS.
10	CLASS supports vastly increased data volumes (compared with current SAA operations) with capacities in the petabyte range.	CLASS is designed to take advantage of emerging technology for data storage (online, near-line, offline, and archived). Data storage strategies such as distributed archiving and HSM are maintained as options. A long-duration resource, CLASS absorbs technology refreshment as new data management concepts are reduced to practice.
11	CLASS evolves over time to achieve its goals.	The immediate goal of CLASS is to support the campaigns providing data in the near future, specifically POES including (DMSP), GOES, NPOESS, NPP, EOS, NEXRAD, and Metop. The first priority of the CLASS Project is to provide the services (including physical data storage and backup) for the flood of data that these campaigns will generate. The CLASS development roadmap identifies sequencing of other capabilities to be implemented over time.

3.2 Strategic Direction

3.2.1 Change Drivers

3.2.1.1 Data Volumes and New Campaigns

Taken from the 2001 Architecture study, the satellite data campaign estimates listed below have not been revisited for this study. These projections were used in planning and acquisition of hardware elements for the first release of CLASS. The in situ campaign figures are current estimates provided by NODC personnel.

Data Category	Est. Annual Growth (TB)	Est. Disk (GB)	Estimated Start Year	Year 1 of Data Start, in TB (compressed)
POES (DMSP)	13.0	107	2002	10
GOES	17.5	144	2002	9
NPP	1,000.0	8,219	2005	800

Data Category	Est. Annual Growth (TB)	Est. Disk (GB)	Estimated Start Year	Year 1 of Data Start, in TB (compressed)
NPOESS	2,000.0	16,438	2010	1,600
NEXRAD	61.5	505	2003	12
NOMADS	95.0	781	2005	76
NOS Side Scan	126.0	1,036	2005	63
EOS (MODIS)	3,000.0	24,658	2004	2,400
Metop	520.0	4,274	2006	14 TB/yr (Uncompressed)

The estimates for Oceanographic archive data include major new archive requirements to be supported through CLASS.

Campaign	Initial year	Archive data volume (TB) per year
Coral Reef Data (Coral reef may include video, which will not be part of CLASS unless it is digitized first by the producer.)	FY 03	10
IOOS System	FY 05	10
SST Reprocessed Data (1-deg resolution)	FY 03	8
Ocean Exploration	FY 03	2 (increasing to >5)
ARGO System	FY 05	1

Additional modeling during the design effort develops estimates for transaction volumes, number of users, search and retrieval load, and other workload quantities.

3.2.1.2 Customer-Centric Ease of Access

Two drivers improve the access to NOAA's data repositories. One is the government-wide focus on providing better citizen service through Internet technology (the Office of Management and Budget's (OMB's) eGovernment initiatives are examples) while reducing the cost of those services. The other driver is the development of new products and expanded uses of environmental data from multiple instruments and collection systems. These more complex products require more sophisticated discovery and selection mechanisms.

3.2.1.3 Advancing Technology

The steady advance of hardware technology for data storage and processing changes the comparative cost of offline vs. online storage. Concurrent advances in data management technology, such as Geospatial Information System- (GIS)-enabled, large-volume databases and data warehouses, expand the possibilities for delivering sophisticated products from a portal.

Faster processing hardware and complex computing architectures (such as large-scale symmetric multiprocessing) and higher bandwidth network capabilities help CLASS in meeting the high-volume requirements of large-array data campaigns. Researchers who take advantage of the new technology are the drivers for CLASS adoption of new capabilities. In particular, such advances increase the demand for a service-based component architecture to simplify the support for new products.

3.2.1.4 Security and Reliability

There are numerous security issues surrounding archive, access, and distribution. Some are general, involving secure management of private information (e.g., user profiles), payment information, access records, and system security and integrity. Others are data stream and collection specific, such as access restrictions and release-date embargoes. CLASS storage and access systems must accommodate federal initiatives on information security and electronic commerce (e.g., digital signatures and authentication).

3.2.2 Major Topics and Issues

To accomplish the CLASS vision detailed above, several issues must be addressed:

a. Building on SAA model to provide support for existing campaigns and information systems

b. Characterizing the customer and user populations (current and expected)

c. Characterizing the information products that fall within the CLASS umbrella

d. Evaluating advanced information technology (e.g., geospatial database systems)

e. Supporting charge-for-data options in an efficient manner

f. Possible integration of "shopping cart" and "charge for data" processes across all NESDIS public information systems

g. Integrating a common user account management capability

h. Supporting non-electronic delivery mechanisms of data.

Each of these is discussed below.

3.2.2.1 Building on SAA Model to Provide Support for Existing Campaigns

The SAA is an efficient and successful system dealing with a small range of data types and providing a somewhat limited user interface capability. It provides effective and reliable input processing, inventory, storage, and retrieval. It supports thousands of users and distributes large quantities of data electronically. It provides data file retrieval and subsetting, subscription services, and bulk-order processing. As the model for large-array dataset stewardship for upcoming satellite campaigns, it represents a solid baseline.

The vision for CLASS goes beyond the current capabilities of SAA. Data collection and analysis campaigns at the NNDCs address a wide variety of observations, often with relatively small volumes of data. These campaigns have developed their own analysis and visualization systems, user communities, and distributions systems (web sites, for the most part) that do not easily fit

into the SAA model. The vision for CLASS must provide value for such existing information systems within a common framework.

Using a one-stop-shop approach for data involves some changes in process and culture, as existing campaigns are successful and effective in their somewhat stove-piped operations. The user interfaces and analytical tools have adapted, often over a period of years, to the requirements and expectations of their distinct user communities. There is some sharing of information (especially metadata) via initiatives including the NOAA Server, but this sharing is less penetrating than envisioned for CLASS. The target architecture and transition plan must provide benefits and incentives for coordination.

A plan for involving and responding to observation campaign management and leadership is essential to achieving the CLASS vision.

3.2.2.2 Characterizing User (Customer) Populations

Before it is possible to define a customer-centric approach to data distribution, the customer base must be identified. The existing customer base can be described by identifying users of existing NOAA/NESDIS information systems. However, a more complete view requires characterizing the users and potential users in terms of expected uses. Since users range from casual browsers to highly focused researchers, a range of customer interfaces must be developed. A variety of browsing and searching tools must be provided. Before these can be defined, a model of information users must be developed.

For the most part, the user base is well characterized. Users added through better access (e.g., educational users at the middle school level) are largely interested in derivative data products such as visualizations and maps. The customer base for raw satellite data is likely to be highly sophisticated and therefore able to navigate complex data systems.

There is a potential for adding users from naïve to sophisticated through the availability of consolidated data products. Information that combines data from two or more disparate campaigns and made available through the CLASS portal may attract a new set of customers.

The planning for CLASS should include a description of the user populations and expectations for new customers as the range and scope of data expands.

3.2.2.3 Characterizing CLASS-Relevant Information Products

In exploring various NOAA, NESDIS, and NNDC web sites, one finds a wide variety of data products, categorizations, browse lists, search tools, and retrieval mechanisms. These include imagery intended for the news media, detailed information on satellite ground tracks and status, and many other items. CLASS primarily addresses the needs of large datasets and high-volume, satellite-generated observation in terms of storage and distribution, but it also addresses (as a portal) data discovery and access for other data collections and derivative products. Policies and decisions are needed to determine what services are provided to different data systems and user groups.

3.2.2.4 Evaluating Advanced Technology for Information Management

The SAA model provides effective storage, file inventory management, and the retrieval of a limited set of data types. However, many existing data types are not included, and the bar is continually being raised for user access and data selection. Mapping software and cross-reference tools will be investigated as additions to existing SAA-based techniques.

3.2.2.5 Supporting Charge-for-Data Options

The SAA currently does not charge for data. Within NESDIS, the NVDC product service and online store provide data products for which fees are charged. These fees are collected through various mechanisms including prepaid purchase accounts, checks, and online credit card interactions. Once NESDIS specifies pricing and payment policies and, potentially, mechanisms to support congressional mandates to recover distribution costs by charging for data, CLASS will adhere to them.

3.2.2.6 Integrating "Shopping Cart" and Payment Mechanisms

The vision for CLASS and sound architectural principles suggests that there should be only one process for "add to shopping cart" and "specify payment" mechanisms. However, many of the interactions with selection and ordering interactions will be implemented in external data discovery systems. The products ordered are delivered through various mechanisms and channels, some of which are not easily centralized. Consolidation of the order and order-fulfillment process involves some flexibility and modularization of the system.

3.2.2.7 Integrating a Common User Account Management Capability

A goal of CLASS is to provide a single entry point for users to discover, order, and get information. A central customer account management system should be an element of the common access (portal) capability (e.g. COAST).

3.2.2.8 Supporting Non-Electronic Delivery Mechanisms for Data

The SAA currently delivers datasets only electronically, many without intervention on either sender or receiver end. Some users, however, require data on transportable media. Many products that come under the purview of CLASS over time will similarly be distributed on physical media. The CLASS inventory and distribution mechanisms must accommodate these transactions.

3.3 Supplier, Customer (User), and System Expectations

The view from outside the system is the end-user view. End users include suppliers of data or information services, data customers, CLASS management, and CLASS operations and maintenance. Each of these groups has requirements for specific services and capabilities.

Suppliers can be classified as:

a. Direct data providers to CLASS

b. Legacy data providers

c. Associated ("independent") data providers

They all have the same basic focus—delivering acquired or derived data to users on request (including subscription services). The role of CLASS is more or less central depending on the classification. Some systems may have greater volumes and variety of products available for distribution and more sophisticated search and analysis capabilities for users, but the essential transaction is the same.

Customers also come in several varieties. **Advanced** users may know exactly what they want and where it is stored (conceptually); they may be getting data that continues a series or provides current information (updates). **Intermediate** users may know what's available and need specific data to support a specific research effort. **Novice** users may simply be exploring what exists and selecting data products based on unique criteria. **Subscription** users provide a standing order to be filled whenever new products are made available in response to a needs profile. **Commercial** users may pay for the products they request. Some users may request non-standard delivery modes (e.g., CD-ROM distribution).

The CLASS oversight, management, operations, and development groups have expectations as well of data integrity and storage. Suppliers provide data products to system storage for archive and backup, keeping some data online, other data "near-line," and all data accessible with moderate delay.

Suppliers expect that the data they provide is available to users on demand. Customers expect to be able to identify and order the products they need and receive them efficiently. Conceptually CLASS is the intermediary that handles non-substantive details like user accounting and tape backup. (The term *non-substantive* is used to describe operational considerations, as opposed to *science data*.)

Since two of the three supplier types have heretofore essentially managed their own user interactions, some overlap or conflict is anticipated between CLASS and those suppliers. For example, a legacy site may be equipped to handle user accounts and subscriptions, delivering new products on a scheduled basis to established customers. CLASS also provides a user interface, customer accounting functions, and event-triggered distribution. The balance between the two must be negotiated. Some campaign archive sites may have complex search and browse capabilities tailored to the specifics of the data managed. CLASS also provides search and browse capabilities. CLASS therefore, should provide similar or better capabilities than those capabilities provided by the suppliers—often by incorporating those supplier capabilities by one mechanism or another.

CLASS provides added value when it supports multiple data types and sources. First and foremost, it must deliver at least the same value as the systems it subsumes and/or supplants. It can provide generalized searching and browsing that covers data from different observation platforms in a common interface. It can support analyses that span multiple data collections. It can provide data options based on multiple collections (i.e., different image resolution, different temperature measurement granularity) to best suit a researcher's needs.

CLASS provides value to NOAA by consolidating functions not specifically related to substantive data issues. To the extent that customer relationship management can be centralized and the overall NOAA data distribution process can be improved, users benefit from a single entry point and access to account information. Suppliers benefit from having the support

functions handled by someone else, allowing them to concentrate on science issues rather than data commerce.

At the same time, there are and should be relationships between customers and observation campaigns. Customer responses and feedback help suppliers refine and extend their product catalogs and improve the usefulness of the data being collected. Suppliers can make customers aware of new or enhanced products. Clearly there must be interchanges between the suppliers and customers—CLASS must (at a minimum) not inhibit such interchanges, and properly should facilitate them.

3.3.1 Supplier Products and Services

Products and services are addressed in two areas: science data and portal functionality. All data suppliers are concerned with the former; independent sites may currently be performing the latter.

This section describes the needs of suppliers and serves as a framework for services that can be provided to suppliers (e.g., by CLASS).

3.3.1.1 Science Data Products and Support

In order to meet the science objectives of observation campaigns and programs, each supplier must accomplish the following tasks (eventually, through future CLASS portal system). The list is focused on data product retrieval and distribution and on science-driven exchanges with users. It is not intended to be a comprehensive list of actions performed by suppliers, and it does not address development of products and supporting software capabilities. The list is limited to data management and science data user support.

a. Archive science data as it is generated
b. Produce basic products (standard datasets and metadata)
c. Produce analytical products (summaries, etc.)
d. Catalog and store (archive) products for retrieval
e. Accept requests for products
f. Deliver products upon request
g. Deliver products for standing orders and subscriptions
h. Provide data-specific browse and search capabilities for users
i. Provide analytical support capabilities for users
j. Exchange information with users about products and potential products
k. Publicize new products and capabilities
l. Provide assistance to users in selecting, retrieving, and using products
m. Deliver special-format products

3.3.1.2 Portal Functionality

Campaigns that use CLASS directly as their data archive, access, and distribution system essentially leave the portal functions to the CLASS portal. Other suppliers with independent user

interfaces generally incorporate these functions into their primary user interface. This segment addresses the user support (e.g., generic search support) that is not specific to data products supplied by a specific campaign.

a. Manage user accounts (create, validate, delete, support)
b. Manage user sessions (authenticate, track)
c. Provide mechanisms for requesting products (including subscriptions)
d. Provide delivery tracking support
e. Provide cost quotation for non-free elements (products)
f. Support payment for products
g. Support feedback and requests for assistance
h. Provide dataset-based search and browse capability.

3.3.2 Customer Expectations

To some degree, customer expectations mirror the supplier capabilities requirements. Customers expect to be able to access the archive, typically via the World Wide Web), and to locate and request information from NOAA's observation programs. With their expectations established by experience with electronic commerce, customers expect to be able to place an order, pay for it (if required), determine order status, and receive the requested products in a reasonable and predictable amount of time. They expect the following:

- To learn about new data products as they become available
- To find a record of previous orders and an ability to place and modify standing orders for products
- To keep their interactions with the archive private, at least at an individual level.

Science data customers have specialized expectations:

- To provide input to the data analysis efforts on observation campaigns
- To influence the campaigns themselves by identifying specific requirements or applications of the potential observation data products
- To be able to provide feedback. They expect to be able to call for help with the data or the distribution mechanism.

3.3.3 System Expectations

The system-level view of CLASS identifies the expectations of planners (operations) and owners (management) for proper operation and effective data storage and dissemination.

3.3.3.1 Management and Oversight Requirements

The expectations of the CLASS management team include operational capability and performance goals and system metrics collection and studying. CLASS must meet some detailed and explicitly quantified requirements for data ingest, archive, and distribution. It must also meet less-well-defined, less-easily-tested requirements for customer satisfaction, operational efficiency, cost avoidance, security, and disaster recovery capability. All of these requirements are addressed in two ways: implementing CLASS to meet performance objectives (subjective as well as objective) and implementing a metrics data collection and studying system to provide a basis for management control.

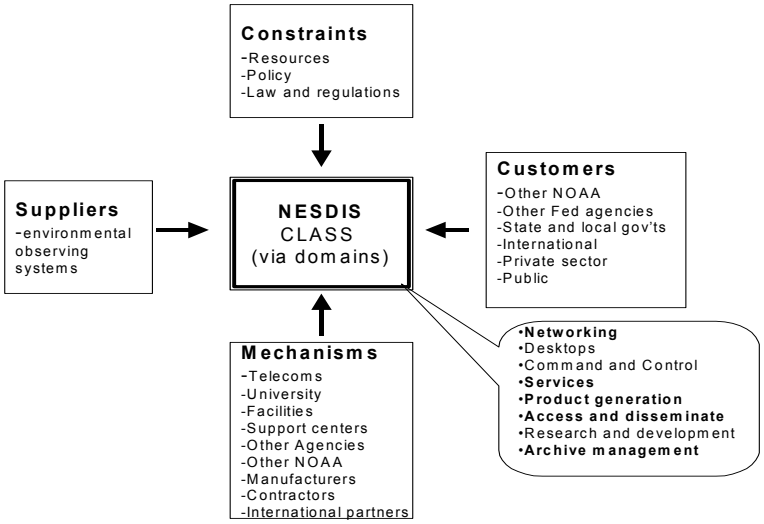
CLASS is expected to provide online access to metrics data (with appropriate access control) to support management activities ranging from long-term planning to responses to emergencies.

3.3.3.2 Operations Expectations

System operators and administrators need the tools and services to support routine operations including data ingest, quality control, hierarchical storage management, backup and restoration, support for users, and system maintenance. One goal of CLASS is to automate as much of the process as is feasible and be cost-effective so that dramatic increases in data volume do not require equivalent increases in operations staff.

3.4 Organizational Context

At the level of NOAA/NESDIS, the context for CLASS involves suppliers, customers, mechanisms, and constraints that affect the ITA domains. CLASS particularly affects the Networking, Services, Access and Disseminate, and Archive Management domains, and somewhat affects the Product-Generation domain. Figure 6, taken from the NESDIS ITA, shows this context.



Business logical context for NESDIS (source: NESDIS ITA)

Figure 6. CLASS Exists in the NESDIS Context

Operationally, CLASS responds to the Suppliers-Customers relationship. Organizationally, CLASS's system architecture business analysis is reflected in the Constraints and Mechanisms.

3.5 Drivers and Use Cases

In the logical architecture, the system elements are defined and validated by mapping to drivers and requirements. Scenarios or use cases provide a convenient approach to characterizing a system. Use cases frequently overlap, as the goal is completeness rather than precision. Requirements can be derived from the drivers and use cases to support system design and development.

The use cases for CLASS (and the associated drivers) are grouped into elements based on the process components that are key to the scenario. Obviously some components will be referenced more than once, and in different groupings.

These scenarios do not attempt to describe detailed hardware selections or software applications, in order that the descriptions remain accurate for the life of CLASS (during which most of the hardware and some of the software will certainly be replaced). Detailed discussions of physical components are presented in the physical architecture section (Section 4).

Scenarios have been developed to describe the user perspective in the following areas:

a. **Data Acquisition**: activities involving delivery of data, of various types, for processing by CLASS
    1. Ingest data (standard workflow for receipt of observation data and metadata)
    2. Ingest independent metadata (not accompanying an observation dataset)
    3. Ingest replacement data (a modification of the standard workflow)
    4. Update metadata (automated process)
    5. Update metadata (web/GUI)

b. **Configuration Changes**: changing the operational structure of CLASS, involving software, hardware, procedural, interface, or standards-based specifications
    1. Add a new data stream to CLASS
    2. Add a new product to CLASS
    3. Add a new external repository
    4. Integrate a supplier-supported software module
    5. Add new top-level metadata
    6. Add or change web content on the portal
    7. Incorporate a new data format standard
    8. Establish an Interface Control Document (ICD) for a new data stream

c. **User Interactions**: exercising system-user interfaces to accomplish various interactive tasks
    1. Establish a new user profile
    2. Modify or delete an existing user profile

3. Create subscription or bulk order
4. Create standard order online
5. Create standard order of Application Protocol Interface (API)
6. Create standard order via CSR
7. Provide feedback to CLASS or observation programs
8. Provide user support such as help screens, FAQs, and access to reference material
9. Provide shopping cart support for external repository

d. **Data Discovery**: specific interactions for the purpose of browsing and/or searching the data collections

1. Explore the catalog and metadata via web interface
2. Explore the catalog and metadata (API)
3. Use advanced discovery tools via web interface
4. Extend search through CLASS portal to external repository

e. **Distribution**: processing sequences designed to deliver requested products to users

1. Process order for electronic delivery
2. Process order for physical delivery
3. Retrieve data from external repository
4. Retrieve data via API

f. **Archive Maintenance**: activities to ensure the continued integrity and availability of archived information in compliance with NARA guidelines and good engineering practice

1. Migrate archive data to new storage system
2. Test and verify archived data

3.5.1 Data Acquisition

Scenario: Ingest Data (Standard Workflow). The data supplier and CLASS have previously established an ICD and/or submission agreement for data transfer defining network channels, file and data formats, processing workflow, and acknowledgment protocol.

The data supplier processes observation data and packages it for ingest by CLASS. The packaged data, in a predefined format, includes the observation data (content) and several types of metadata: descriptive of the observation, technical details of data capture and processing, administrative information about the packaging, and discovery information such as browse imagery. The supplier either places the package in a retrieval location for CLASS to scan or pushes the package to CLASS.

CLASS accepts or retrieves the package and performs the agreed-upon processing such as extraction of metadata and transmission-integrity QA. An acknowledgement is sent to the provider, identifying the package and any processing anomalies.

CLASS creates access records (inventory-level data) for the data package and all distinct components. Modified metadata (e.g., with CLASS access information) is written to the Archive, Access, and Distribution metadata repository. A subset of the inventory-level metadata could be written to the NOAA discovery portal.

CLASS writes some or all of the data to permanent archive (based on data-stream profile) and to online storage. Some data may be posted to a CLASS GIS database for programmatic access and advanced discovery processing.

If requested, an "archive written" acknowledgement will be sent to the data supplier (e.g., to facilitate safe purging of storage).

CLASS compares the inventory-level metadata with existing subscription orders and triggers creation of delivery order(s) as appropriate for the newly archived data.

Scenario: Ingest Independent Metadata. The data supplier and CLASS have previously established an ICD and/or submission agreement for metadata transfer regarding observation data that is maintained outside of CLASS. The ICD and/or submission agreement defines the format and content of the metadata as well as data transfer channels, processing workflow, and acknowledgement protocol.

The data supplier creates inventory-level access and discovery metadata for observation data. The metadata includes reference information (e.g., Universal Registry Locator (URL)) to support access and cross-references as appropriate to other discovery information. The metadata is packaged and either placed in a retrieval area or pushed to CLASS.

CLASS retrieves the package, extracts the data, and performs transmission-integrity QA. The information is written to the NOAA discovery portal. Web page information is updated, if necessary, through automated updates. An acknowledgement is provided to the sender.

Scenario: Ingest Replacement Data. In accordance with the data-stream profile, when CLASS retrieves or receives a dataset whose unambiguous designation matches one previously received, CLASS processes the dataset. Typically, CLASS evaluates the QA score of the new dataset, compares it with the equivalent value for the archived copy, and either replaces the previous dataset (erasing it), discards the new dataset, or keeps both versions in the archive. Generally, the quality score of the new dataset must be significantly better than the old dataset in order to trigger a replacement action.

Note: In an automated process, the decisions as to whether or not old and new versions of reprocessed data should both be archived are made on a data-stream profile level rather than on a dataset-by-dataset basis.

Scenario: Update Metadata (Automated Process). With appropriate security, data providers can update the metadata records maintained for the data they have provided. Newly received data is automatically reflected in the inventory-level discovery database, but directory-level metadata may not always be updated due to the complexity of the relationships. Data suppliers can provide automated updates to predefined elements of metadata through programmatic interfaces.

Data providers send update transactions to CLASS. The transactions are verified for authentication and internal consistency, applied to the metadata database(s), and logged. Update confirmations are provided programmatically and, if requested, by other channels (e.g., email).

Scenario: Update Metadata (Web/GUI). With appropriate security, data providers can update relevant metadata records through a web-based interface. All of the CLASS-maintained metadata is available for modification by authorized users. Some types of changes will be delayed pending review (e.g., changing the name or top-level description of a data campaign); these generally involve the elements that are displayed at the top level of the observation data hierarchy.

For changes requiring review, the automated workflow process directs the change records to the cognizant personnel to assure timely implementation of approved changes.

3.5.2 Configuration Changes

Scenario: Add a New Data Stream to CLASS. The data manager of an environmental data repository wants to arrange for CLASS to serve as the archive and distribution system for the observation data. The process involves defining the data stream, getting approval for the effort, specifying the data transfer channels and volumes, creating entries to support data discovery, and initiating the support.

Note: integrating a major new data system (e.g., a satellite campaign) is a major undertaking that typically involves system development and engineering. Those major initiatives are treated separately but include most or all of the steps identified here. This scenario involves addition of a small-scale observation data stream to CLASS.

The data manager creates a formal request for CLASS support services through a negotiated ICD or submission agreement. The request includes information of the customer base, the data volume (existing and projected), processing and discovery requirements, science applications of the data, funding sources, and other pertinent information. The DAB and the CPMT review the request. This scenario assumes that the request is approved.

The data manager defines the data stream in accordance with CLASS standards, specifying the data communications channels, ingest processing, metadata records, data hierarchy, and so on. This information includes an ICD (discussed below) and other processing support details that comprise the data-stream profile. The profile includes the top-level metadata entry for the NOAA discovery portal, framework for directory-level metadata, and descriptions of product interrelationships, web-page entries, and inventory-level metadata definitions. The profile also includes the processing rules for accepting or rejecting datasets, storage rules (e.g., how long a data component should remain online under various circumstances), and processing priority agreements. The profile specifies how ordering, delivery, and any necessary payment are to be addressed within CLASS. The profile may include specialized software to be hosted on CLASS systems to perform QA or discovery processing, or on-the-fly product generation. The process for adding such software is discussed below. Decisions on adding such software are made by CLASS (e.g., the CPMT).

Some data streams may have a backlog of data to be loaded into the CLASS archive in advance of (or concurrently with) the continuing data stream. Procedures and schedules for such catch-up loading are negotiated with CLASS operations.

The data provider negotiates the agreement with CLASS, establishes the data transfer channel (typically an FTP interchange), verifies any changes needed in the CLASS-maintained web sites, and initiates transfer. Any bulk loading activities are also initiated on the schedule as agreed.

Scenario: Adding a New Product. A data manager or research team has developed a new derived product based on one or more existing data products. For example, multiple satellite orbit data and in situ observations are combined in a mosaic. The producers want to make the products available through CLASS. Users of the product express a desire for subscription orders of products as new source data becomes available. The request for CLASS is to archive and publish products as suppliers generate them.

Adding a new product is similar to adding a new data stream. Rules for data storage may differ. Certain data products may be destined only for time-limited storage rather than permanent archive in CLASS, where long-term storage for these kinds of products is hosted somewhere else.

The data manager creates a request to the DAB and CPMT as above. A decision may be made that specific product datasets that are ordered are archived; other products are made available online for negotiated periods of time and then purged. (Presumably any derived product based on archive data can be recreated, if necessary, at least for some period of time.)

Scenario: Adding a New External Repository. A data manager of an environmental data collection (not necessarily within NOAA) wants the data collection to be publicized and accessible through CLASS. The actual data, as well as content-specific data discovery tools, remain on the external repository. CLASS includes top-level and directory-level metadata with links to the owning information system.

CLASS can provide a variety of services to an external repository, from a simple listing in the high-level metadata directory to tightly coupled access and distribution services. The data provider develops a plan for using CLASS and negotiates a shared responsibility agreement. The CPMT makes a formal request for review. (The DAB may not be involved because the data is archived outside of CLASS.) The request describes the services required.

Once the request is approved, a data-stream profile is created so that the repository is accurately studied in the metadata repository. Procedures are established for metadata updates. Other services (such as links to advanced data discovery tools, user directory services, or shopping-cart ordering integration) are engineered to support the repository.

Scenario: Integrating a Supplier-Supported Module. The data manager for an existing CLASS-supported data stream has developed a new orbit scan data subsetting that may be of general use. The software tool provides improved performance in creating geospatial subsets of satellite data and is applicable to many data streams. The tool has been developed in accordance with CLASS software development standards and is intended to be compatible with the CLASS infrastructure.

The providers develop or submit a CLASS Change Request (CR) for review by the CLASS Configuration Control Board (CCB). At the direction of the CCB, the CLASS System Engineering Team (SET) evaluates the request, performing tests and impact estimates. The change is approved, and the providers and CLASS personnel integrate, test, document, and deploy the new capability.

At some stage in the process, the tool is publicized to other data managers and may be adopted for use by multiple data streams. The data-stream profiles are changed accordingly as part of the deployment process.

Scenario: Adding New Top-Level Metadata. A new data stream has been added to CLASS (whether internal or external) and the NOAA data discovery portal must be updated to reflect the changes.

As part of the process of adding a new data stream, a new top-level metadata record is created to serve as the entry point. The top-level information is reflected in the portal web site in several ways. In addition to the search-based access, the data stream is publicized in the portal "what's new" section, and a separate web page may be created for detailed information or it could be pushed to registered users.

Organizational impacts include identifying the content owner for information about the new data stream (typically the data manager, but large campaigns may have separate points of contact) and establishing roles and responsibilities for metadata management. Training in the use of tools for metadata maintenance may also be appropriate.

Scenario: Add or Modify Web Content on the Portal. A data supplier changes the information presented to customers about a specific data collection or stream (e.g., announcing a new data product).

An approved data manager (DM) logs in to the discovery portal content management system and is authenticated with edit rights to the content reflecting his or her data collection. Using any suitable tools [dependent on the selected content management system], the DM creates new content or edits existing content. A new GIF or JPEG image may be inserted or a news brief created. The DM may add a reference to an upcoming symposium, tagging it with a "display until" date. When done, the DM posts the new content and logs out of the content management system.

Depending on established practice, the new/modified content is either published to the portal or routed to a web page review process for approval.

Scenario: Incorporating a New Data Format Standard. In order to support Coral Reef Information Systems video data, a new version of the MPEG standard must be added to the CLASS suite of tools. The data manager identifies the requirement and tools needed to support the standard and proposes it to CLASS.

The data manager develops a request to CLASS identifying the data format, the applicable standard(s), the volumes of data, and the necessary processing to be performed on CLASS infrastructure. Processing may include compression/decompression, playback, extraction of metadata, verification of integrity, subsetting, and creation of browse images. Compatibility with existing data standards (if any) supported by CLASS will be documented. Software appropriate to CLASS will be identified. Other aspects (e.g., data volumes of the specific data stream) are addressed elsewhere in the set of scenarios.

The CPMT and possibly the DAB review the proposal. If there are competing or equivalent standards, there may be a review by a designated panel to select the best choice. The NESDIS CCB may review the selection as well, as it may affect the NESDIS ITA. (If the new format is already defined in the ITA, this step may not be necessary.) Implementation of appropriate tools follows the scenario for integrating a new module into CLASS.

Scenario: Establishing an ICD and/or Submission Agreement for New Data Stream. The CPMT and DAB have approved addition of a new data stream. The details of information exchange must be defined and documented. The process results in an ICD.

CLASS and the cognizant data manager negotiate the communication channels (e.g., FTP drop-box procedures and addresses), expected data volumes, error conditions and actions, ingest processing expectations, data formats, metadata types and structure, and any other specific requirements. The ICD negotiation process ensures that expectations and commitments have been clearly understood, and the ICD becomes part of the formal CLASS documentation. Once complete, the ICD is placed under configuration control.

3.5.3 User Interactions

Scenario: Establish a User Profile. An unregistered user creates a user profile for the purpose of ordering data products.

The user, visiting the NOAA Data Discover (DD) Portal, clicks on the "register" button. At the registration screen, the user enters an email address and standard demographic data. A search is made through the CLASS user directory for a match.

If a match is found (possibly because the user had previously registered on a repository system whose user base was folded into the CLASS directory), the user is asked to review and verify the information maintained. If not, a new entry is created for the user.

This user is interested in products for which fees may be charged and elects to provide a credit card number for the sake of convenience. The credit card number entry will be encrypted and will never be displayed to the user. Other options are available, such as establishing a cash account with NOAA or identifying an organization (Government agency, University, etc.) account previously established; those options require interaction with a customer service representative. Users interested only in free products skip the payment options process.

This user is interested in information that may be restricted. Accordingly, the user provides additional information that permits verification of status and against current "watch lists" as defined (e.g. the name of a college or university and reference information). Other information may be collected for the user profile, such as preferences for web server startup pages.

The user profile is created, and a confirming email is sent to the specified address with an initial password.

Scenario: Modify or Delete an Existing User Profile. An already-registered user changes some element(s) of the user profile or requests that the profile be deleted.

The user logs on to the discovery portal and clicks the link for "update your registration." A data entry/edit form is displayed with current data (exclusive of sensitive information such as credit card numbers). The user can delete the registration (requiring a confirmation) or change the information. Some changes, such as credit card information, require a secure link and may require some additional form of authentication. The user reviews and commits the changes.

An email confirmation is created and mailed. If the user changed his or her email address, a confirmation is sent to both addresses.

The order management system is searched for any active transactions (in-process orders, pending payments). The orders are adjusted, if necessary, to reflect the changes.

The changes are logged for audit purposes.

Scenario: Create Subscription/Bulk Order. A regular (already-registered) customer requests an extended order either to retrieve a large volume of data or to retrieve any new data that meets a certain specification.

The first step involves the customer creating the specifications of the request. This can be done online with a web browser or with a narrative description. The online approach may be based on a search: the search arguments can be saved and used as part of a subscription or bulk order. The order may be submitted through the web, emailed, faxed, surface mailed, or placed telephonically.

The second step involves review of the order by a CSR. The actual order is analyzed for data volume and impact. For bulk orders, a delivery schedule is created. After reviewing the conditions and expectations with the customer, the CSR enters the order.

Scenario: Create Online Order. The customer has identified one or more data products for retrieval and delivery. The distribution includes both electronic and physical channels.

Through data discovery mechanisms and filtering, the user has created a record (a "shopping cart") of several data products to be delivered. Each item (or collection) was identified through a search or browse process and selected for delivery (an "add to shopping cart" button or checkbox for each item). Once the customer has selected all the items, the "review your order" selection is made and the summary list of items is presented for review. The user can drop any items from the tentative order, resume the data discovery process, or cancel the activity all together. The expected prices of any priced items are displayed and a preliminary total cost is also displayed.

The user selects "place the order" and the order is recorded. An email confirmation of the order is sent to the recorded email for this customer.

Scenario: Create Standard Order (API). The customer has established a relationship with CLASS and an approved automated interface to the order management system. Based on any of several data discovery mechanisms (e.g., email notification of data), the customer's computer system places an order. For example, a derivative data producer may receive notification whenever a specified combination of observation data products are ingested into CLASS; the data products are then ordered as input to the next-stage derivative (which may in turn be automatically provided to CLASS).

Establishing the channel involves a secure handshake between CLASS and the customer's system. The automated interaction does not require an email notification as the verification is provided directly. However, the delivery notification may still use email because of the delay involved. The conditions are established as part of the specific interface protocol.

With automated notification and ordering, derivative products can be created and made available with minimal delay.

Scenario: Create a Standard Order. Orders involving special conditions are placed with a customer service representative. The process is essentially the same as for subscription and bulk orders.

Scenario: Provide Feedback to CLASS or Observation Programs. The user provides comments or questions either to the portal system in general or to a specific data supplier.

A user clicks the "feedback" function button on a discovery portal web page. A comment screen is raised, and the user adds comments. The user has choices including whether to provide a name and contact information, whether to direct the comments to a specific organization, and so on. The user submits the feedback.

Depending on the context (i.e., where in the discovery system the feedback was triggered) and the choices made, the comment is logged and directed to the point of contact for the selected activity. CLASS personnel also log the comment for review.

Scenario: Provide "Shopping Cart" Support for External Repository. An external repository within the NESDIS domain uses the CLASS order management system. The repository may or may not use CLASS for data delivery or data archiving.

The repository data discovery system includes links to the CLASS order management system to create, modify, price, and place orders. The CLASS user directory provides user information for ordering. The repository generates XML-formatted requests to CLASS to initiate an order, place items in the order record, and perform any modifications the user requests. Once the order is complete and approved by the user, the repository software instructs the CLASS (Order Management System (OMS) to place and close the order.

If the data products are not archived in CLASS, the order-item details include the necessary detailed information to direct the external repository to deliver the requested data—either to CLASS for subsequent distribution or directly to the user.

3.5.4 Data Discovery

Scenario: Explore the Catalog/Metadata (Web). A customer accesses the NOAA DD Portal page and explores the data collections available. Several different sub-scenarios involve different purposes and levels of user sophistication.

Research efforts and some commercial activities (e.g., weather forecasting) commonly involve one or more data streams over an extended period of time. Researchers generally search by discipline in data collections already explored and look for data with specific temporal or geospatial attributes. These customers may use broader search capabilities to identify related data (e.g., searching for drought and/or flooding indications in concert with precipitation data. The user interface for these customers should provide ready access to the directory-level metadata to identify specific datasets. (Note that once a research project is underway, researchers are likely to establish subscription orders for specific data streams and not use the NOAA DD Portal.)

Some research efforts focus on particular environmental phenomena (e.g., drought) and may involve exploration of possible interconnections in data. For these customers, the top-level metadata should provide adequate descriptions of the data collections to identify commonalties. This would include links from observation data to derived products, which are described in the portal. Once a set of collections has been identified, further searches by time and location should

identify dataset-level information. Users are able to review the attributes of selected datasets, including quick-look browse images if available, to refine the selection.

Other users of the data discovery system, such as educational users and the general public, probably search for specific products in a particular time and place. A comparison of NOAA server access logs in different regions of the United States shows that customers are likely to select states in their own region. This suggests that customers want information about their own locations (e.g., local weather satellite images). Vacationers may look for ocean temperatures, current time, at specific coastal locations. The discovery interface allows users to choose by place names (gazetting) and event names (e.g., Hurricane Hugo), and to select products by instructive names ("beach water temperatures") as well as by more rigorous terms ("coastal bathythermography"). Users can also filter their choices by cost, format, and delivery mechanisms to further narrow the search. In many cases, the associated browse images for data may be the desired product.

In each case, the customer has options to provide feedback to the CLASS program and to pose questions or ask for assistance. The portal will provide guidance and explanations to simplify the discovery process.

Scenario: Explore Catalog/Metadata (API). A research program, once underway, may choose to place subscription orders for continuing data streams or to perform periodic automated searches of the metadata repository. Researchers perform searches interactively to fine-tune the selection parameters and then save the query terms for repeat use. The researchers can write their own search-and-retrieval programs to interact with the portal web server using a published API. This allows an automated update process for research data and may be coupled to an API-based ordering or data retrieval mechanism.

Scenario: Use Advanced Discovery Tools (Web). Advanced data discovery mechanisms are available in two forms: those hosted on the NOAA DD Portal and those hosted on data-stream-specific systems. This scenario addresses the second option; the requirements for the first mode of operation have not been specified.

As an example, a customer may create a selection list of satellite data in the discovery portal and then apply further criteria (such as, "cloud-free over Detroit"). The Discovery portal does not support this content-specific search, but the data-stream information system may have the desired capability. In this case, the Discovery portal establishes a link to the data-stream web server and transfers the selection list (possibly in the form of the relevant search criteria); the user then interacts with the content-aware web site to refine the search and make any desired retrieval selections.

The content-aware server may be part of CLASS (e.g., the SAA data discovery system) or part of an external repository. If the server makes use of the CLASS user directory service, the transition may be essentially transparent to the user. Otherwise, the user may need to re-authenticate to the new server.

Scenario: Extend Discovery Process to an External Repository. The user chooses to use discovery resources provided as a pass-through from CLASS.

While in a discovery mode, the user decides to drill down further than the level of detail that is provided by the generic discovery tools on the CLASS portal. The user has selected a data product that is stored in relational form on an external repository and wishes to set parameters

unique to that data product. For this scenario, the product is a gridded dataset of two or more data sources that can be used to generate display images based on a complex data selection set of parameters. The external repository provides a graphical interface for the user to select parameters and adjust thresholds, viewing quick-look samples of the product. (Think of Photoshop picture quality controls for a reference.)

The CLASS portal does not provide the GUI interface to the relational data repository; so the user is referred to the external repository. The link to the repository web site invokes an interface with the necessary data to return the user to the CLASS discovery portal. The user makes a selection that defines a specific data product (a unique image) and chooses "add to shopping cart."

The external repository generates the product (or schedules its generation), provides a unique identifier, sends an XML-encoded message to the CLASS portal with the order information, and directs the user's session back to the previously established session. The new order item has been added, and the user can proceed to place the order, cancel the order, or continue shopping.

3.5.5 Distribution

Scenario: Process Order for Electronic Delivery. An order has been generated and placed. Any required payment information is available and is approved. The distribution system locates the data and provides it to the user.

The order consists of order items, each one describing a specific dataset or on-the-fly data product. Each order item is treated independently. Each delivery is transmitted separately. The system aggregates the items placed over some short period of time and sends an email notification. Since some items may take longer to retrieve than others, this process allows some data to be received quickly even when other data is delayed. Requests for data that are online are usually filled within minutes, while data in the archive may take hours due to physical retrieval times and the volume of orders. On-the-fly data products, such as subsets of orbit-based satellite datasets, also take additional time.

Generally, the order is filled from the archive site that is the primary site for the data requested. However, either archive site can provide data to be placed in either FTP site.

Once retrieved and/or processed, the data products are copied to an FTP staging area. The items transferred are listed in an automatically generated email message that is sent to the recorded address of the customer. The customer is requested to retrieve the data within a specified period of time (e.g., 72 hours). This email is in addition to the order confirmation email sent when the order is placed. If necessary, more email messages are sent as more data products are placed for retrieval.

The customer (or an automated process) receives the notification and uses anonymous FTP to retrieve the data. Once the information is retrieved, the product can be purged from the FTP area. The timing on this purging depends on the data traffic. If there are charges associated with any of the products, the customer's account is billed accordingly.

During this period, the user can inquire via the web or through an API as to the status of the order. The status identifies all files placed for retrieval as well as any order items still pending.

If an error occurs in the retrieval (or if the user simply does not retrieve the datasets), the customer can use the order information process to request that the information be posted again.

At any time before the order is complete, the customer can cancel the order (or whatever remains of it.)

If the data are restricted, the CLASS distribution system provides data encryption and electronic signature services to ensure that only authorized users can retrieve and process restricted data.

Scenario: Process Order for Physical Delivery. The sequence of events in this scenario is largely as above (electronic distribution) with added elements of physical selection and shipment, and inventory control of discrete items.

The order management process generates pick and pack instructions and labels for the physical media staff located at NCDC. If the order requires standard products, the products are picked from inventory, packed, and shipped. For data products that are not already written to tape or CD, the required datasets are retrieved and staged for copying. The physical copies are made and then packed and shipped. In either instance, the process is recorded by the order management system, and the inventory (of products and/or media) is updated.

An email notification of shipment is sent to the customer. The order status is changed to "shipped." If there is a problem with the order (e.g., a tape is unreadable), the customer can contact a CSR to arrange a repeat shipment or cancellation of the order with refund as appropriate.

Scenario: Retrieve Data From External Repository. The process is largely as above. The order contains a product that does not reside in the CLASS archive or online cache. However, the data manager for the repository in which the product is found has arranged to have CLASS do product distribution. CLASS sends a request (probably XML-formatted) to the external repository. The repository processes the request (with CLASS as the user) and provides the product via the agreed-upon channel (probably an FTP site). The repository information system notifies CLASS programmatically that the product is available (linking it to the request); CLASS retrieves the product and treats it as if it had been retrieved from the CLASS archive.

Scenario: Retrieve Data Via API. In addition to the web-based automated retrieval capability, CLASS supports access to data products through the use of APIs. There are two approaches to automated retrieval: one creates orders (mimicking the web-based interface); the other provides direct access to the CLASS data stores (with appropriate security).

The web-simulation API is a general mechanism particularly suited to coupling with automated directory searching. When a search mechanism discovers that a desired data product is available, an order is created, effectively through the web interface.

The direct-access mechanism provides a tighter coupling to the CLASS data management system. Products that are cached in the CLASS product database can be searched, manipulated, and retrieved through a web-service-based database connection. This process provides the customer with the full power of the database engine (including text search, SQL selection, and geospatial operations) to select and pre-process data. In particular, the interface provides a capability for integrating information from different products and data streams for specific purposes. The product generation is performed on the customer's system, but a significant fraction of online data is available for these data fusion operations. Customers can essentially create new products from stored and managed data.

The direct-access mechanism requires a higher level of authentication and security than other interfaces. CLASS database security provisions ensure that the managed data cannot be changed, that only authorized users can access the data, and only the data for which they have been granted access rights.

3.5.6 Archive Maintenance

Scenario: Migrate Archived Data. Copy archived datasets from one medium to another (e.g., a newer tape format) while maintaining integrity and all access metadata.

Details of this operation are developed during system design based on specific file and data management characteristics of the old and new media systems.

Scenario: Test and Verify Archived Data. Perform routine or as-needed testing and recertification to ensure the integrity and accessibility of data.

System logs of data written and of retrieval actions (and errors) are used in an automated analyzer to trigger periodic verifications of data. Test programs are written to perform statistical analysis of archived datasets, with care to include all dataset formats and a mix of recent and older data. Verification standards are set.

Details of the operational test process will be developed during system design.

3.5.7 Sample Scenario for End-to-End Data Flow

As an illustration of the scenario development and application, the following composite scenario was developed to show the process from receipt of data by CLASS, triggering a standing order, to delivery of the new content to a customer.

#	Scenario Step	Comments
1	Dataset is written to inbound FTP site.	Supplier (e.g., an observation program or campaign) writes a dataset to the CLASS inbound FTP site.
2	Receipt is detected.	CLASS continually monitors the FTP site, detects the submission (which may be either scheduled or ad hoc), and flags the information.
3	Workflow package is initiated.	A workflow package identifying the files, time, date, any inventory information, and parameters from the data-stream profile is loaded into the job queue.
4	Ingest processing initiated.	When resources are available, the received data is analyzed and specific processing steps are scheduled. Datasets are moved to staging locations for processing.
5	QA verification occurs.	Integrity, data denial, and readability of the data is checked, along with any data-stream-specific checks such as duplicate file, comparative quality (based on supplied flags) or other tests.

#	Scenario Step	Comments
6	Inventory record created.	If the dataset does not pass the QA step, a negative acknowledgement is sent and no further processing occurs; in this scenario, the data is entered into the CLASS operational inventory.
7	Acknowledgement is sent.	Successful (or unsuccessful) ingest status report is sent back to the supplier (could be placed in the FTP area for retrieval by the supplier).
8	Local archive is written.	Any permanent records are written to the local archive.
9	Remote archive is written.	Permanent records are written to the remote archive, where no ingest processing occurs. Database records (e.g., Ops Inventory) are synchronized.
10	Content is formatted for Content DB.	If a GIS-DB format has been defined for this type of data, the dataset is reformatted and written to the content DB where it is available for content-specific search, manipulations, and retrieval.
11	Metadata is written to Metadata Database (MDDB).	Metadata is extracted from the file and constructed (perhaps with data-stream-level information maintained within CLASS) and written to the metadata DB. The ops Inventory is updated to include the metadata references.
12	Standing orders checked; order item created.	The acknowledgement process includes internal publishing of the data availability; one step in the publishing is a check of standing orders (subscriptions, block data) that may trigger a distribution.
13	User account is verified.	An active order is identified, an order item is created, and the user's profile is checked to ensure that the order is valid and any necessary payment information is current.
14	Data is packaged for distribution.	The data is retrieved (from cache, archive, or by Content DB extract) and packaged for delivery. A workflow package is generated to accomplish the order.
15	Data package posted to outbound FTP site.	The packaged data is loaded into the outbound FTP area.
16	Customer is notified by email.	The customer is notified by email that the data is available and will be kept in the FTP area for a limited time. Other notification mechanisms are available, defined as part of the standing order.
17	User retrieves package.	
18	Retrieval is detected.	CLASS monitors FTP transactions and generates a workflow to close out the order and the overall workflow.
19	Order item is closed out.	Any necessary billing or other transactions are completed, and the order item is closed and recorded. (The standing order remains open.)
20	Workflow package is closed.	The workflow triggered by the receipt of data is closed.

This sample workflow is used in a description of the target architecture elements in section 4.

3.6 Architecture Views

This section discusses different aspects of the architecture that are critical to understanding the approach taken by CLASS.

3.6.1 Features and Design Elements

This section addresses salient elements of the architecture that are reflected in the physical and design architectures. These elements are presented here in order to establish a basis for discussion of process and data organization with common terminology.

3.6.1.1 Observation Data Hierarchy

Environmental observation data is provided to CLASS in the form of datasets, packaged aggregates of data (e.g., files) in agreed-upon formats. Different operations within CLASS operate on data at different levels of detail. Figure 7 shows the overall classification of data reflected in CLASS.

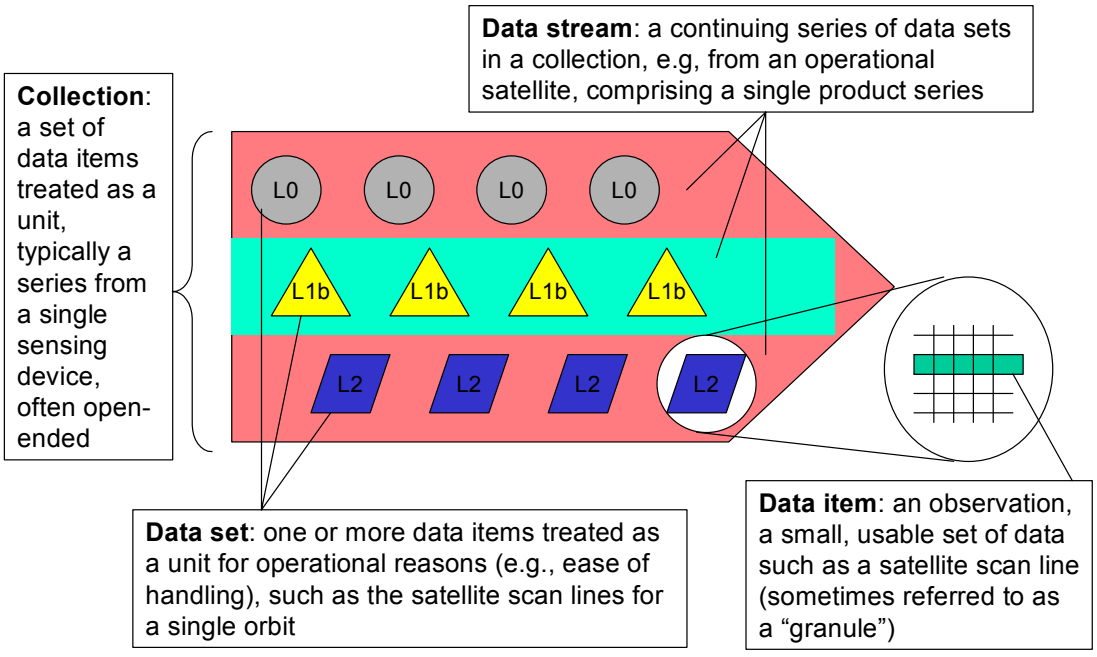


Figure 7. Conceptual Structures of Observation Data in CLASS

The smallest item of data with which CLASS is concerned is a data item, typically a single-sensor reading or coordinated set of readings. A single temperature reading from a buoy sensor might be treated as a data item. The granularity of data archived and distributed by CLASS is an attribute defined in the ICD for a specific data stream.

Data items are grouped for convenience and processing performance into datasets, which are aggregates of data items in a specific format. A dataset (usually a file, although composites can exist) is the typical package of data provided to CLASS.

A data product is any, defined, separately accessible collection of data, typically a dataset but (in the case of RDBMS-based information) potentially any usable processed package of data. A series of datasets (e.g., a time series) for a single sensor or product specification comprises a data stream. A set of related data streams, such as Level 0 and processed derived products, are identified as a data collection.

The distinctions are important for data storage and data discovery process descriptions.

3.6.1.2 Metadata Classification

CLASS manages, archives, and distributes data such as observation data and descriptive information about the observation data or metadata. Both types of information are essential to the effective use of the environmental record. Within CLASS are different aspects and applications of metadata. Figure 8 shows the primary classification of metadata in this context.

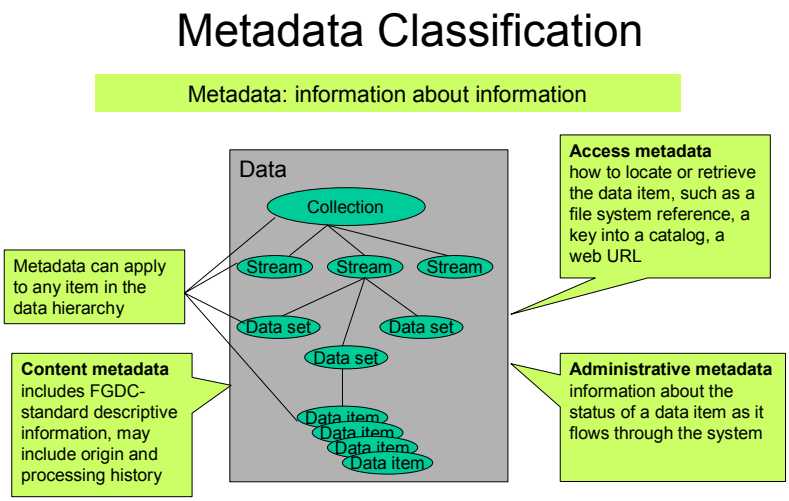


Figure 8. CLASS Metadata Classifications by Application

CLASS metadata includes content metadata, typically generated by the data provider and supplied with the observation data, access metadata, information used in retrieving datasets or data items, and administrative metadata developed and used in the process of acquiring, archiving, publishing, and distributing data products.

As noted in the figure, metadata can describe data at any level of the information hierarchy from complete collections down to individual granules.

3.6.1.3 Data Discovery Hierarchy

Using metadata at different levels of the observation data hierarchy provides a comparable hierarchy of data discovery. The discovery function can be distributed among different operating

elements of NOAA's repository system. Figure 9 shows the allocation of metadata discovery for different observation data levels.

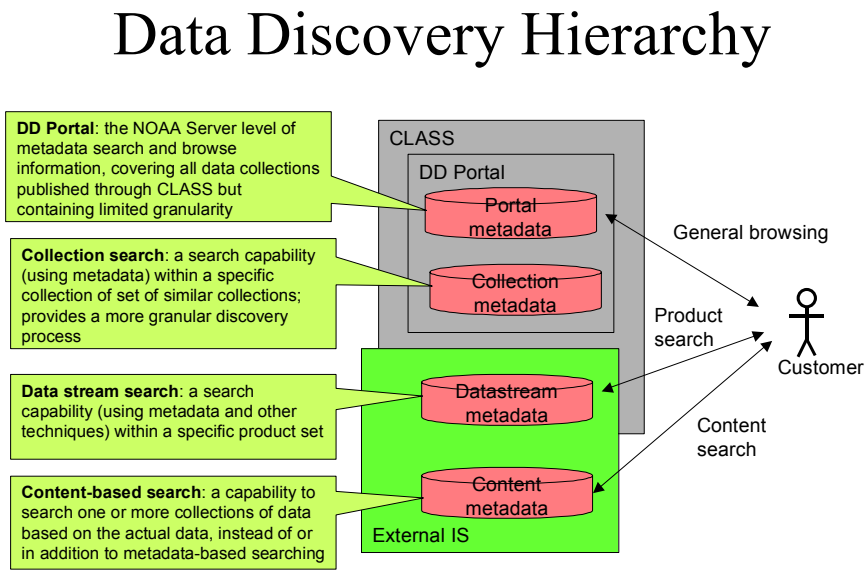


Figure 9. Users Can Drill Down Into the Data Discovery Hierarchy

The figure shows the overlap in allocated responsibility between CLASS and external information systems. The capabilities overlap to provide the most effective service to users without excessively loading the CLASS portal.

Note that the metadata described in Figure 9 above is essentially content metadata that describes different levels in observation data hierarchy. That is, data-stream metadata is content metadata about the data stream but not about the data-item level of metadata.

3.6.1.4 Permanent and Transient Data

Because it serves as both a portal and an archive, CLASS is designed to manage both permanent and transient data. Permanent or long-term record storage is provided using NARA-compliant archive technology and processes. Other information, such as un-validated data products, multiple versions of products, discovery products (e.g., browse images), or topical data (e.g., water temperatures along beaches during vacation season) may be kept online while there is a demand and then deleted. CLASS provides retention schedules and rules to ensure optimum benefit to customers while making effective use of online storage.

Oversight boards establish principles and guidelines and evaluate specific requests for retention periods. Data-stream-specific guidelines are established during the development of ICDs.

3.6.1.5 Distributed Redundant Archive

CLASS provides total duplication of archives while providing load-balancing of data processing activities. The archives and associated metadata are synchronized between two (or more) archive

sites. To enable either instance of CLASS to support inquiries about any CLASS-maintained data, the appropriate discovery metadata are duplicated along with the archived data. Users have a one-stop shop at either archive while the processing load is shared between the Suitland and Asheville sites. Figure 10 shows the distributed architecture.

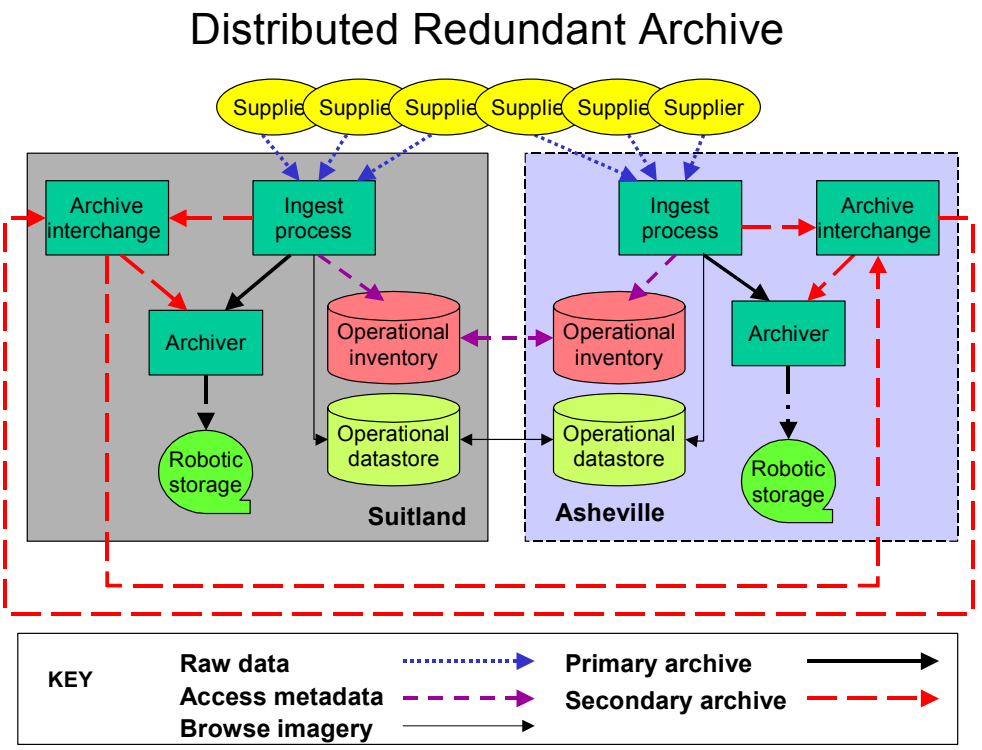


Figure 10. Integrated Archive With Operational Redundancy

The archive interchange components provide assured transport and verification between the two CLASS instances. The operational inventory stores access metadata, while the operational data store contains discovery metadata such as browse imagery. The synchronization between the two instances permits allocation of users to one site or another without restricting discovery and retrieval capabilities.

3.6.1.6 Integration With External Repositories

Acting as a portal, CLASS makes use of external resources to meet customer needs. At the same time, CLASS provides services to external repositories (within NESDIS) to minimize duplication of effort and provide a consistent interactive behavior to users. The integration involves definition of common services triggered with XML-based messaging.

One element of the CLASS vision is the minimization of stovepipe system development for storage, access, and distribution of environmental data. Research teams have developed dozens or hundreds of information repositories, each with its own archive, ingest, discovery, user base, delivery system, web server, and administrative overhead. In the baseline system environment, a research team has no other way to store and publish its data. CLASS is not planned as the home to all these data systems and (because of process-intensive unique data operations, if for no other

reason) cannot serve as the global home. Current direction is that CLASS is not intended to provide archive space for these data streams.

CLASS can, however, provide common services to these research-based repositories to integrate them into the NESDIS CLASS-centered operations model. If more external repositories provide up-to-date metadata for discovery purposes, integrate their access and retrieval mechanisms, and share a user base, there will be a powerful argument for integrated data support.

The functional integration methods described here need not be the only such integration. Stovepipe systems are anachronistic, and given the choice, most research teams would probably not build their entire repositories as independent systems.

3.6.1.6.1 Discovery Pass-Through

External repositories may provide more content-specific and detailed discovery capabilities than the CLASS portal supports. To provide users with the greater specificity of the repository-based searching, CLASS can pass through a search request to the external system and receive a selectable list of products as a result. Selections from that search result may be passed back to the CLASS repository for retrieval and processing. If the selected products do not reside on the CLASS data archive or operational data store, then CLASS, acting as a portal support system, can pass the request to the external systems. Figure 11 illustrates this pass-through process.

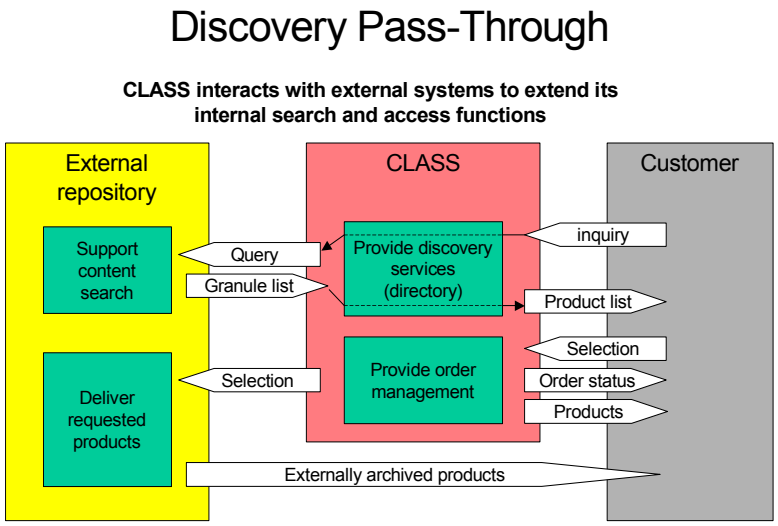


Figure 11. Supplementing CLASS Discoveries With External Repository Services

External repository systems that were selected to be part of the CLASS portal system may require modification to achieve this integration. CLASS defines the interface protocols for distributed discovery processes and product delivery requests.

3.6.1.6.2 Common Services Support

A number of common services are used in any archive, access, and distribution system. Some functions are very content specific, but many (i.e. user services) are common services that can be built upon shared data structures and integrated service frameworks. CLASS provides such

common services as user profile (customer account) management, authentication, and data discovery. Figure 12 shows CLASS in the "back-end" services role for a NESDIS external repository.

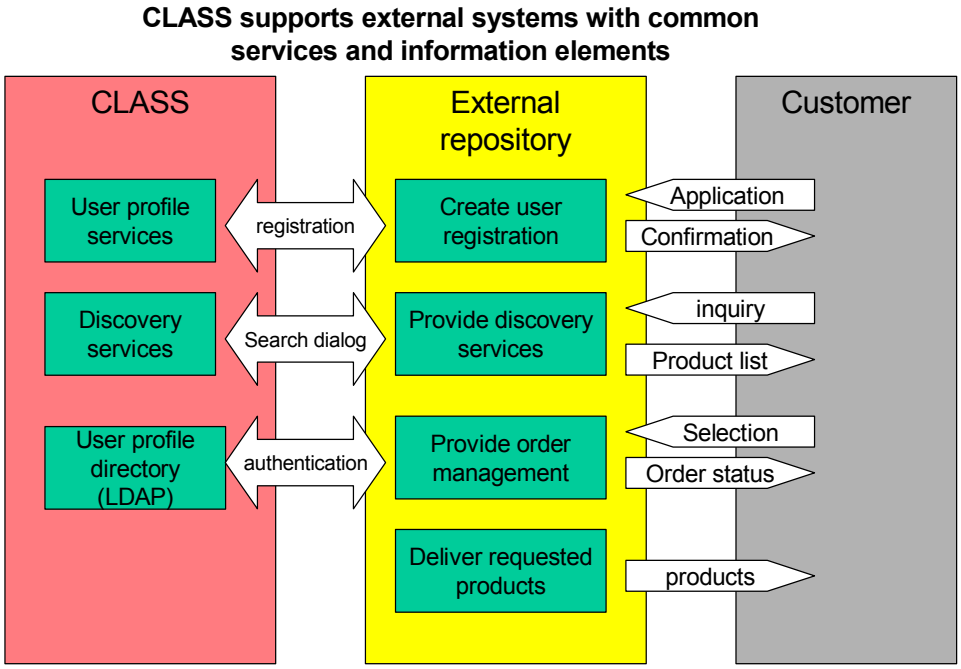


Figure 12. CLASS Services Facilitate Repository Engineering

The right side of the figure shows the interaction between customer and repository; the left side shows some common services supported by CLASS. Because the external repositories do not have to develop and provide these services internally, researchers can focus on improving the science value of their systems. Because common data and protocols are used among the repositories, customers do not have to deal with as many disparate logins and interfaces.

3.6.1.6.3 Order and Distribution Support

In addition to the data services noted above, CLASS provides operational services such as order management and data distribution, further reducing the development and maintenance load on the content-specific, research-driven repositories. Figure 13 shows the integration between CLASS and an external repository for order management (e.g., "shopping cart" functionality) and product distribution. CLASS defines the protocols and messaging contents to support this integration.

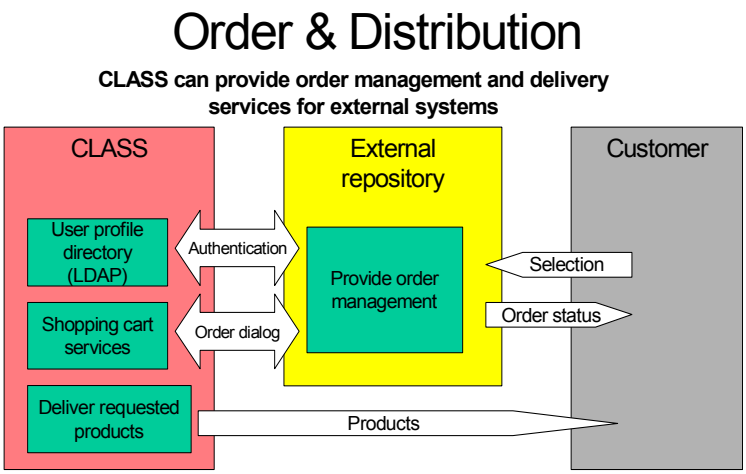


Figure 13. Support for Common Ordering and Distribution Functionality

These integrated order and delivery capabilities permit customers to select and order products from different repositories (including CLASS) in a single transaction, thereby minimizing the "seams" between the different systems.

3.6.1.7 GIS-Enabled Online Data Store

CLASS provides, as part of its online storage capability, a Database Management System (DBMS)-based product repository with GIS capabilities appropriate to the products. Not all data products are suited to DBMS residence. There is limited value to including data products as "blobs" (binary large objects) when the HSM file system does a perfectly good job in that respect. Many products, however, can be searched and modified with the GIS functionality available in modern DBMSs. The specific products to be "database'd" will be negotiated with suppliers.

Once in the DBMS, the products will be accessible to programmatic access through web-service protocols, giving researchers improved access to data.

3.6.2 Process View

The CLASS logical architecture includes a discussion of the business processes needed in the system to address business drivers. These processes can be described in generic process flows and descriptions of the process hierarchy.

A process flow diagram presents process activities in a matrix of roles and sequence information that characterizes a process in a visual manner. The process flow shown in Figure 14 includes the generic end-to-end set of activities that take received observation data through to a customer's order and receipt.

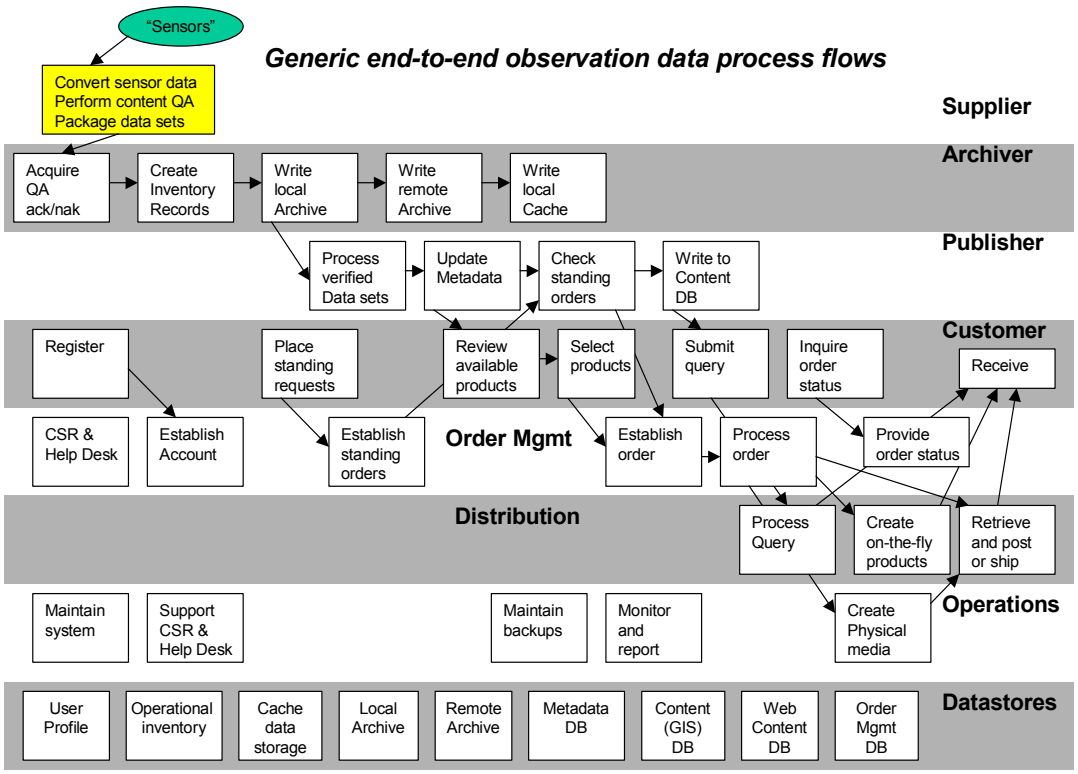


Figure 14. Generic Observation Data Flow

The Archiver role involves the activities surrounding writing data to online and nearline storage. The Publisher role publicizes the availability of the received data. The Customer registers, explores, orders, and receives data. Order Management keeps track of new and standing orders and performs the workflow associated with ordering. The Distribution role includes the retrieval, packaging (including dynamic product generation, if required), and distribution of products.

The Operations role involves the routine management and customer-driven support activities. The Data Stores layer (not precisely a role) serves as documentation of the functionally different data storage requirements of the activities in the process.

Additional detailed and specific process flows can be developed during the design phase to clarify detailed design specifications.

3.6.3 Data Organization and Flow

3.6.3.1 NOAA/NESDIS Environmental Data

The focus of CLASS is the environmental observation data and related data products collected through NOAA observation campaigns and research activities. CLASS is not directly concerned with the scientific characteristics of the data. Stewardship encompasses safeguarding and distributing the data but not dealing with substantive content. Stewardship includes data discovery and delivery, and therefore includes data classifications, formats, attributes, locations,

security and access considerations, and support tools such as data viewers. CLASS is particularly concerned with data volumes and collection rates to ensure that adequate resources are available.

3.6.3.1.1 Data Formats

Data formats used by CLASS are determined through ICD and/or submission agreement with data suppliers, discussions with customers, and evolution of IT standards. Low-level formats are developed in accordance with industry best practices and NARA guidelines. Higher-level formats are often observations data specific and are subject to negotiation. As a general rule, information is stored in ASCII text format or self-describing data formats such as netCDF and Hierarchical Data Format (HDF), or commonly accepted image and multimedia formats such as GIF, JPEG, and MPEG.

3.6.3.1.2 Metadata

Content metadata is primarily maintained in the FGDC-extensible standard as text information. For operational purposes, it may also be stored in relational, GIS-enabled database systems from which standard XML formats can be published. Internal metadata (access and administrative) are developed for optimum performance during system design.

As noted earlier, *content metadata* describes the substantive information in a file or similar data aggregate. Content metadata, typically captured in the CLASS environment in FGDC standard, can include the origin of the data, processing history, observation target, and similar descriptive information. Most *data discovery* processes operate either on metadata or on the data directly. *Access metadata* is file-oriented (or record-oriented) and describes locations of data aggregates. *Inventory information* is primarily access metadata. *Administrative metadata* is process-management information that describes what processes are performed or scheduled for a data aggregate.

Data discovery encompasses searching and browsing through the observation data and the content metadata to locate usable data items or composites. (A subset of an AVHRR orbit collection is actually a composite of individual scan lines.) Effective discovery requires access to both data and metadata.

3.6.3.1.3 Data Standards

Data standards applying to CLASS data include HDF, Net CDF, FDGC, XML, HTML, and Extensible HyperText Markup Language (XHTML).

3.6.3.2 Data Management Concept

The instruments used to make observations are outside the scope of this architecture study. Similarly the specific data types and formats of observation data are not addressed here. CLASS is mainly concerned with managing the data itself, while the uses to which data is put by customers is outside the scope, as well. The general classes of data, however, such as time-series observations, large-array data (raster images), network observations, data products, and metadata, are matters of interest. The scope of the architecture extends from the receipt of data (or data products) into a collection to the distribution of raw or processed data to customers.

3.6.3.2.1 Supply Chain Functionality

The basic elements of the forward flow model differ from campaign to campaign, but in general support the following functionality.

a. **Ingest**: receive or actively acquire campaign-generated data, perform data quality checks, convert data to standard formats, extract metadata for cataloging and inventory, and move data to appropriate storage locations.

b. **Catalog and Inventory**: process, organize, and store file metadata to support data retrieval and data discovery.

c. **Data Processing**: generate product files to support customer interaction (e.g., visualization data); perform analytic processing for explicit requests; this complements the processing performed by campaigns prior to the ingest step.

d. **Archiving**: manage the storage of data; for CLASS this involves hierarchical (near-line) storage systems in multiple, redundant locations, but other non-CLASS repositories may use different tactics.

e. **Data Discovery**: provide multiple ways of finding, subsetting, combining, visualizing, and analyzing data as an aid to data product selection and request.

f. **Customer Management**: manage user accounts (including access privileges, subscriptions, and payment mechanisms).

g. **User Interfaces**: includes both online (e.g., web based) and system-to-system interfaces to provide access to the archives.

h. **Order Management**: provide ordering mechanisms (online, subscription, etc.) based on selections made in the data discovery process, tying product selection to the product distribution process.

i. **Distribute Data**: (also known as order fulfillment); retrieve, assemble, process and package (as necessary), and place products for customer retrieval.

This model captures the core capture-archive-deliver functionality of stewardship. It does not reflect the essential interactions between information customers and the observation campaigns. The detailed nature of this core functionality is based on science goals and the user communities that exist around those goals.

3.6.3.2.2 Feedback and News: Supplier-Customer Interactions

Numerous established user communities have close and productive ties to NNDC data projects. Through these ties, NNDC can propose and test new products and delivery formats, and users can identify needs and expectations. This collaborative relationship has been generally effective at increasing the value of observation data. The stewardship architecture concept recognized this relationship and adds IT-based feedback mechanisms.

Figure 15 identifies the forward and reverse information channels that are integral to the information system model. It does not reflect the rich context in collaboration that encompasses the system, as that context is outside the scope of this architecture.

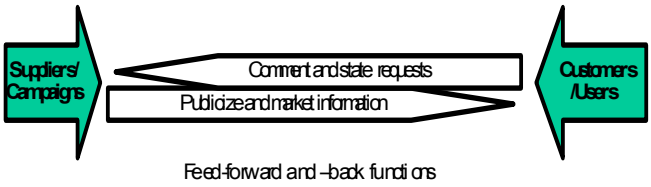


Figure 15. Supplier-Customer Interaction Channels

The stewardship systems exist in a context of continuing collaboration. The information systems provide news (e.g., about new data products or analysis results) from the suppliers to customers. Active and passive channels (e.g., direct feedback and web click stream logs) provide information from customers to suppliers.

3.6.3.2.3 Management, Direction, and Operations

Information systems exist to accomplish explicit goals, but they must be constructed, managed, operated, reviewed, and occasionally redirected. The CLASS stewardship model addresses the systems context with the crosscutting functions illustrated in Figure 16.

Figure 16. Stewardship Management, Direction, and Operations

1) **Manage and Report**: track the process of receiving, storing, and distributing data, providing a basis for future planning and providing feedback to data suppliers.
2) **Manage Distributed Information Processing**: Allocate and track tasks across multiple locations, servers, and storage devices to achieve redundancy and load-balancing.
3) **Provide Information Security**: secure and protect information assets from accidental or malicious damage and from unauthorized access or modification.

4) **Operate**: carry out all routine activities to ensure that services are provided as expected and that system capacities are not exceeded.
5) **Maintain**: provide routine corrections and refreshment to ensure effective operations.
6) **Implement, Integrate, Expand**: support the continued evolution and expansion of the system to meet the needs of new campaigns.

For smaller data collections, all of these functions may be the province of a single person or a small group. For larger systems (such as the current SAA and CLASS), the overall architecture must explicitly define and provide organizationally for these functions.

3.6.3.2.4 Business Architecture: Conceptual View

Putting these elements into a single figure provides a basis for discussion of the architecture elements of the current and target stewardship capabilities. Figure 17 identifies the functions that need to be performed without tying them to a specific physical implementation.

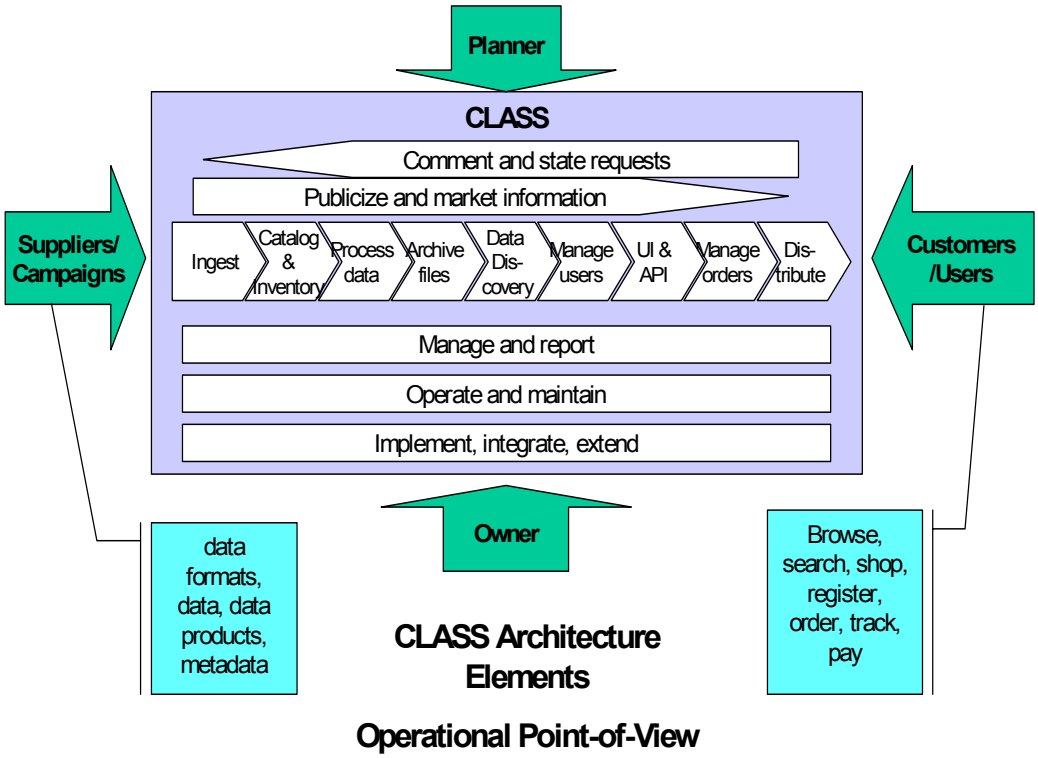


Figure 17. Business Architecture Conceptual View

3.6.3.3 Data Model

CLASS operates on the following types of data. Detailed data models are developed in the design phase of development.

CLASS Data Type
FTP inbound staging data
FTP delivery staging data
Job-step sequence data
Process workflow data
Workflow templates
Distributed-archive management data
Data-element inventory
Access metadata
Administrative metadata
Content metadata
User profiles
Physical-item inventory
Order-item data
Order data
Discovery (browse) images
Web content
Observation data
Operations and system logs
Email records
Documentation and help data
Web-service publishing data

3.6.4 Applications

Applications will be required to perform the following functions:

a. Ingest observation data and metadata
   1) Acquire datasets
      - Pull from external system
      - Accept via push FTP
      - Acknowledge receipt
   2) Perform transmission QA

3) Perform data-stream QA
4) Process datasets
   - Compile and extract metadata
   - Reformat observation data for online access
   - Generate browse support products
5) Create inventory records

b. Store data
1) Write to archive and verify
2) Place in online storage
3) Update inventory
4) Acknowledge receipt
5) Synchronize archives
6) Synchronize data stores
7) Manage archives
   - Perform test and verification
   - Perform data migration
8) Backup and recovery

c. Publish data availability
1) Check standing orders
2) Update DD portal metadata
3) Update data stream and content metadata
4) Update web content

d. Support data discovery
1) Search metadata
2) Search directory
3) Search content
4) Search visualization
5) Search data interchange
6) Search analysis
7) Search external repository
8) Provide selection of data products for distribution
9) Support API access to discovery services

e. Support ordering information
1) Create orders

2) Create order items
3) Provide tracking information on orders
4) Fill orders with physical or electronic distribution
5) Support API access to ordering services

f. Create process-management scripts for specific jobs
g. Manage workflow
h. Monitor system activity and performance
i. Study on system activity.

The applications supporting these processes are described in the physical and design architectures.

3.6.5 Technology

CLASS relies on software and hardware technology to support the applications identified herein. The selection of technology involves assessment of commercially available components in the context of open systems and published standards. As CLASS is developed, the selections of technology are made based on best value at the time of implementation.

3.6.5.1 Software System Elements

Software elements of CLASS that represent technology choices include:

a. LDAP user directory service data store and application (e.g., Netscape Directory Server)
b. GIS-enabled Content (cache) database (e.g., Informix RDBMS with Geospatial DataBlade module)
c. Operations data RDBMS (e.g., Informix, Oracle, DB2)
d. Order management system application suite and data store (e.g., Oracle 9i with Oracle Financials)
e. FTP server including secure FTP
f. Web servers (e.g., Apache)
g. Hierarchical storage management system (e.g. SAM-FS, HPSS)
h. Workflow engine
i. Web-content management system (e.g., Interwoven)
j. Web portal server (e.g., Vignette)
k. Application server (e.g., IBM Websphere, Sun Solaris One, Tomcat)
l. Web services framework (e.g., J2EE).

3.6.5.2 Hardware Technology

Hardware elements to be incorporated into the system design (some will be provided through NESDIS network support) include:

a. Portal server
b. Application server
c. Web servers
d. Operations servers
e. HSM storage system (e.g., NAS and SAN systems) for terabytes of storage
f. Robotic tape library for archive for petabytes of storage
g. Network infrastructure including
    1) Routers
    2) Firewalls
    3) Data lines
h. Test environments
i. Development environments.

3.6.6 Infrastructure

The supporting infrastructure for CLASS development, operations, and maintenance includes development tools, test tools, configuration management software, issue-tracking database support, and operational procedures for thorough testing and certification. An operational system, CLASS can be changed only after verifying that the changes do not disrupt operations.

4. CLASS PHYSICAL ARCHITECTURE

The physical architecture describes the components and interrelationships needed to support the required functionality. The baseline architecture describes the starting point for development (in this case, CLASS Release 1). The target architecture describes a system that is capable of meeting all requirements. The transition plan in Section 6 describes a sequence of development that can accomplish the target architecture.

Section 4.1 presents the baseline system, section 4.2 presents the target system architecture, and section 4.3 presents a gap analysis between the baseline architecture and the target architecture.

4.1 Baseline

The current stewardship architecture includes four major operational elements (or categories of elements): the Satellite Active Archive, the NVDS order management system, the various NNDC data repositories, and the NOAA server metadata description and access system. The importance of separating the business architecture from the physical architecture is immediately obvious, as each of these components operates in one or more distinct physical environments using different servers, networks, and software to accomplish some of the same basic goals.

Figure 18 shows these components with primary capabilities overlaid on the basic stewardship model, identifying the functions that are shared among elements and the functional gaps in each element. For the independent repositories, the figure is generic: each repository needs to be analyzed to show how it maps to the model.

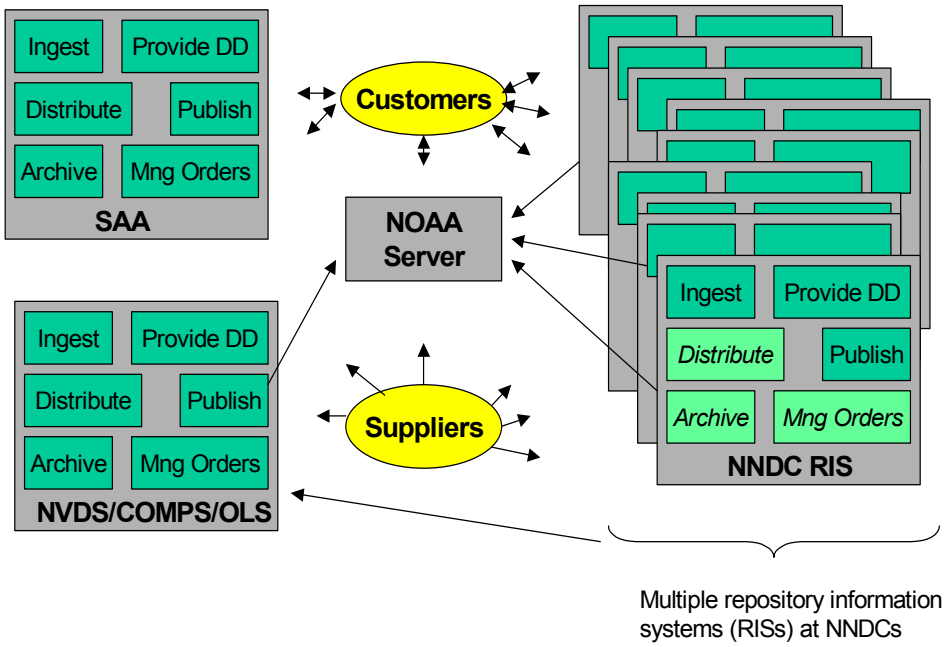


Figure 18. Baseline Architecture Elements

4.1.1 Satellite Active Archive

The most robust and mature of the systems that comprise the CLASS baseline, SAA has been providing large-array data management for years, archiving and distributing large volumes of data. It focuses on receiving, storing, and efficiently delivering datasets as requested by customers through online and operator-controlled mechanisms. It performs relatively little processing and provides basic information discovery functionality or value-added services. SAA manages several different streams of data and data products but does very little in the way of integrating those data types. The SAA does not include any cost-recovery mechanism. It excels at its primary function and provides a suitable base for additional satellite data campaigns.

CLASS Release 1 is an extension of the SAA, re-hosted on new hardware, using a new robotic tape system, and including more "portal" functionality. It also provides redundant, synchronized storage with data archive systems in Suitland, MD and Asheville, NC. It includes management, development, and operations organizational capabilities and features some limited news and feedback capability.

The SAA is largely a "pure IT" system, with little direct customer interaction beyond the distribution of data and help desk and little supplier interaction except in terms of file formats, metadata, quality measures, and operations. The broader CLASS (including R1) governance model does address supplier and customer issues as they relate to data stewardship.

4.1.2 NVDS/Order Management System

The NOAA Virtual Data System OMS component (including COMPS, OLS, COAST) provides a mechanism for users to create orders with selections of specific products, to place those orders, provide payment information, track the status of orders, and cancel unfilled orders or order elements (line items). Hosted at NCDC, OMS supports access to NOAA products from many different sources. It provides a customer database that can include account information for repeated purchases. The existing OMS capability (in COMPS) includes a web interface with electronic delivery but primarily serves as a tool for customer service representatives to use in answering mail and telephone orders. COMPS is currently being replaced by an Oracle financial system.

The OMS interacts with a product inventory system that provides descriptions of all data products available for order in various media (e.g., CD-ROM). The inventory system includes prices for products, with policies that offer modified prices to customers based on customer classification and NESDIS policy. The OMS interacts with an order fulfillment system that includes picking, labeling, and shipping products to fill orders. The OMS provides some of the functionality required in CLASS but not designed in the SAA.

4.1.3 External Information Systems

NOAA research provides a wide range of data products and tools to add value to environmental data. These research activities distribute data through a variety of channels including web access, FTP file transfers, programmatic interfaces, and the NCDC-based OMS. Each of these information systems—data repositories, access and distribution systems, data processing, data discovery and analysis support, and customer base—shares some of the functional characteristics of the SAA. Each may provide some subset of the generic processing flow described earlier.

Typically these information systems have specialized online interfaces tailored to the characteristics of the research, the data (e.g., time series measurements), and the user community. Some of the systems generate metadata for use with the NOAA Server, supporting general access to the broader environmental research and education community.

4.1.4 NOAA Server Metadata Compendium

Descriptions of data products held by NOAA observation campaigns and research are collected and served to the public via NOAA's metadata system, NOAA Server. Each participating organization generates data about its data in accordance with the FGDC standard and provides it to the NOAA server system maintained in Silver Spring, MD.

The NOAA server is the current portal providing access to campaign or collection-level metadata with directory-level browse and search data discovery capabilities. For more detailed discovery or products selection, the portal redirects the user to one of the online repositories or the NOAA online store component of NVDS.

4.2 Target Architecture

The architecture for CLASS is described in terms of the software and functional elements that accomplish the business requirements, the technology infrastructure (hardware and systems software at a generic level) that supports the processes and the organizational framework in which the functionality occurs. The technology architecture is based on a reference framework that supports allocation of functionality to specific layers, which can in turn be allocated to physical components.

4.2.1 Operational Concepts

The target architecture for CLASS is based on a reference architecture using a layered model that has been common in IT systems for decades. The reference model addresses the primary operational activities by grouping functions into *layers* of similar functionality and system requirements. A common conceptual structure, the layer model groups elements that are tightly integrated with each other in layers and separates elements that may have less tightly coupled interfaces. The layers defined for the CLASS architecture are:

a. Presentation (formatting of information for presentation over various media channels)

b. Application (system-to-user support elements, including portal services)

c. Integration (system-to-system communications support components)

d. Services (internal data-processing components and data management)

e. Data (storage and storage management)

f. Workflow and business process management (system control and integration)

Each layer includes components that provide layer-specific functionality, the layer-specific interface technology (sometimes called middleware), and any data elements specific to effective system performance. The basis for allocating functions to elements includes consistency, commonality of purpose and of application type, and issues of efficiency.

For example, the data layer includes the primary data stores, but many applications and components include object-specific data. Such data stores can be implemented in any layer to ensure effective operations of the system and avoid translation or communications bottlenecks.

4.2.1.1 Presentation Layer

The presentation layer provides the formatting and packaging of information provided to users, including web presentation, operations console support, study generation, and targeting of information for specialized media such as Personal Digital Assistant (PDA), email, or Voice-over-internet. The presentation layer consists of channel servers such as web servers and print servers. This layer also includes support applications such as Section 508 support accommodation, dynamic browse image creation, map drill-down, and other specialized display and print functions. The presentation layer is built on standard protocols such as HTML.

Internet and intranet communications connect via appropriate servers through the presentation layer.

The basic rule for including functionality in this layer is that the content is provided to applications (e.g., a web server) and then formatted or otherwise processed at user direction (e.g., "redisplay in printer-friendly format") without recourse to the providing application. For example, zooming in on an over-sampled digital image may be performed in the presentation layer without re-reading the image from a retrieval application.

4.2.1.2 Application Layer

The content presented through the presentation layer is generated largely in the application layer, a collection of system-to-user applications and services. The standard J2EE framework is used to integrate these applications, which are accessed through Internet channels or by the workflow process of service invocation. Applications include the web portal capability, data discovery services, user authentication, and content-processing tools such as data product subsetting and dynamic product generation.

Built on standard interfaces such as J2EE, the application layer encompasses web services standards such as SOAP and UDDI.

4.2.1.3 Integration Layer

Interfaces between the CLASS instance and other systems are mediated through the system integration layer. This layer supports system-to-system interfaces including FTP receipt of data and FTP delivery of products. Other interfaces include XML-based system interactions and the synchronization process between the Asheville and Suitland CLASS archives.

4.2.1.4 Services Layer

The high-volume transactions of CLASS, including data acquisition and ingest processing, archiving, and data store loading are accomplished in the services and data layers. The services layer includes the internal system functional elements that perform the primary processing data flow of observation data management. The database applications such as user account management, metadata storage, and delivery packaging of products are allocated to this layer.

Interfaces among applications are enabled through custom workflow elements, standards-based interfaces, or commercial middleware. Adapters are created as needed (or acquired from COTS vendors) to connect the data in these applications to other elements of CLASS. Because of the high data volumes involved, most of the observation data transfers are mediated through custom workflow that leverages server file systems and hierarchical storage management. Specifically, a workflow package describes required processing and identifies the files to be operated on, rather than pipe large files through messaging queues.

4.2.1.5 Data Layer

The data storage elements of CLASS are considered extensions of the services layer. CLASS combines high-volume data processing with large file sizes, making the data layer a high-visibility component of the layered model. The data layer includes the hardware and software components of the hierarchical storage management systems including the robotics tape systems, network-based storage, and databases.

4.2.1.6 Workflow Layer

Although the term "layer" is not entirely appropriate, the workflow component of the architecture is a distinct and integral piece of the architecture model. The processes that complement the use cases (scenarios) are orchestrated with workflow threads and business process logic. The workflow layer includes the scripts that govern the flow of data and activation of processes in response to external events and internal operations.

The workflow layer includes the various job control and process routing mechanisms that define CLASS. It also includes the operator control and process management components of the system that supports changes to the process. It includes the process data that supports operational activities and serves as the basis for performance metrics and statistical studying. Any triggering activity that crosses layers, and many that operate within a layer, are managed by elements of the workflow layer.

4.2.2 Operational Software Elements

The framework for the CLASS layered model (e.g., the layer definitions and interaction protocols) is based on the e3 architecture framework reference model developed by Computer Sciences Corporation for enterprise integration architectures. The framework supports system development and integration where multiple components and independent systems are to be connected in a flexible and maintainable fashion. The framework prescribes a layered model (with flexibility in layer definition) using a services, portal, and workflow approach to operational integration. The e3 reference architecture is applicable to CLASS because of the variety of system interfaces, both internally and within NESDIS, and the need for internet- and portal-based customer support.

e3, which stands for "enabling extended enterprises," is a service mark of Computer Sciences Corporation.

Figure 19 shows the software elements of the CLASS layered model expressed in terms of the e3 framework.

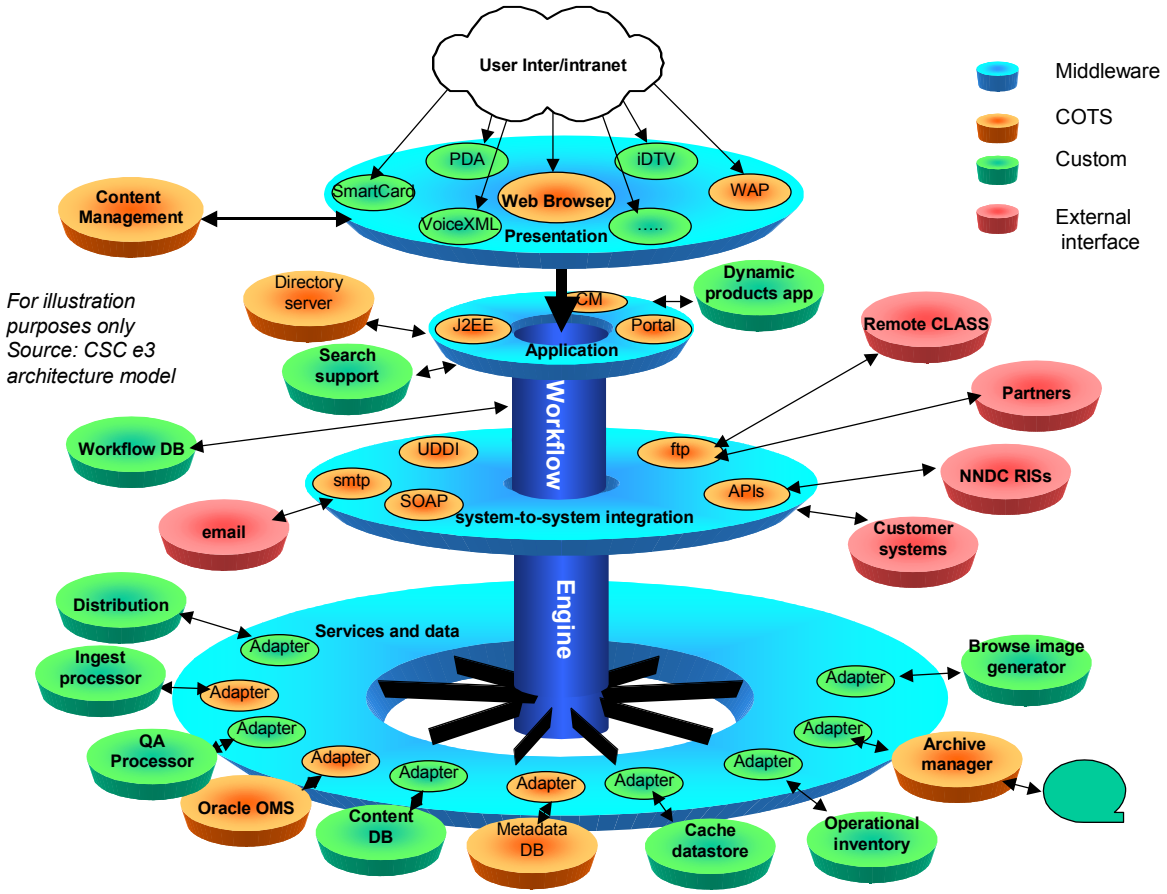


Figure 19. CLASS Layered Architecture Model

Each of the layers and components is discussed in the following sections.

4.2.2.1 Presentation Layer

Initially the presentation layer consists of a web server and associated static web content. Dynamic web content will be managed and provided in the application layer. Special functions needed for Section 508 compliance are included in this layer, possibly as Java servlets or other server-based components. As CLASS evolves, additional presentation layer logic will be incorporated into the capabilities of the system. The presentation layer also includes the operator interfaces and systems management consoles for management of the underlying computer systems.

User interactions with the system may trigger workflow elements or create transactions involving applications such as user authentication. As more communications channels are added, such as streaming video and audio, web pages targeted to personal communications devices (e.g., internet-connected PDAs), and virtual reality media, the presentation layer provides the abstraction from the applications and services layers.

ID	Component	Description	Interfaces
1.1	Web servers	Provides direct interaction with web browsers, publishing static content and linking to services layer for comprehensive and dynamic services; There will be multiple web servers throughout CLASS	HTML, HTTL, XML Internet, Intranet
1.2	Operations consoles	Network and system management interface	
1.3	Study production	Formats and prints content from study generator functions	
1.4	Web content database	Display management for published web content	Web servers Portal Web content manager
1.5	VOIX	Internet-based voice response system	Web content DB
1.6	Retargeting systems	Channel controllers to reformat HTML, XHTML content for various media such as WAP, iDTV, PDA/mobile devices	Web content DB Web services
1.7	Section 508 support	Tools to reformat content to comply with requirements	

4.2.2.2 Application Layer

The application layer contains the server-to-user logic that responds to interactive processing requests. Hosted on one or more application servers, components at this layer perform portal functions such as personalization, user registration and authentication, data discovery, product selection and ordering, order tracking, and user feedback. Also at this level, probably on an internal-use-only application server, are the internal management interactive support capabilities such as configuration management, issue tracking, workflow management, system metrics analysis and studying, and status studying.

Dynamic processes such as on-the-fly product generation (e.g., composite images) and dataset subsetting are performed in this layer. The principle is that temporary products are created in the applications layer; back-end product generation is performed in the services layer.

Applications are invoked directly by interactive processes (e.g., a “sign-in” element in a web page that triggers the user authentication application) or by workflows triggered by such interactions. For example, a new user registration process may trigger related applications (such as writing to the accounting order management system, requesting payment information, or sending email acknowledgements) based on scripting in the workflow system that reflects the business processes of CLASS.

Services that are provided to other systems or layers are defined and published through the integration layer.

ID	Component	Description	Interfaces
2.1	Data discovery portal	Web portal application for public interface and data discovery (e.g., Vignette portal)	Web servers Applications
2.2	Web content manager	Management system for creation, editing, version control, approval, publication of content via web sites and the DD portal	Web servers Web content DB
2.3	Metadata search engine	Collection and data-stream-level search for content	Metadata repository Item order service
2.4	Directory Browse engine	High-level selection and browse tool for ObsData	Metadata repository Item order service
2.5	Feedback support	Mechanism for user feedback at various levels of CLASS	
2.6	Metadata maintenance	User-driven update mechanism for metadata	Metadata repository Web servers
2.7	Item order service	Create an item order from search selection	OMS
2.8	Order tracking service	Determine status of existing order	OMS
2.9	User registration service	Create a new user profile or modify/delete an existing profile	User account DB
2.10	Online dynamic product generation	Create modified products (typically, display images) from ObsData based on predefined data-stream-specific capabilities	
2.11	Online documentation and help	Context-sensitive help and online documentation support	Web content DB Web servers
2.12	System monitoring and management	SNMP-based management and studying service	Operations consoles Workflow metrics System indicators
2.13	Services directory	Published services specifications (e.g., UDDI directory)	

4.2.2.3 Integration Layer

System-to-system interactions are mediated through the integration layer. The process that scans external FTP sites for newly provided data is an interface layer component. The inbound FTP site for data receipt and the outbound FTP site for product delivery are in this layer. The external API support is processed in this layer, providing an authentication process for system interactions. Activities in this layer typically trigger workflow scripts to accomplish activities.

Archive synchronization is accomplished in this layer.

Automated management of the interface systems is also part of this layer. For example, the purging process that deletes files from FTP locations after they have been retrieved (or after some timeout) is an interface layer function.

Services defined in the application or services layers that are intended for web access (web services) are published in this layer, in which the services directory is maintained.

Elements in this layer are typically standards-based (FTP, XML, SOAP, UDDI, etc.).

ID	Component	Description	Interfaces
3.1	Inbound FTP area manager	Monitors inbound FTP operations, generates workflow triggers, purges staging area, maintains links among related files, provides acknowledgements	Ingest process Workflow
3.2	External FTP site retrieval process	Monitors FTP locations of data suppliers for new content, retrieves files, generates workflow triggers, maintains links among related files, provides acknowledgements	Ingest process workflow
3.3	Outbound FTP area manager	Places distribution files in staging area, monitors retrievals, purges area	Distribution process workflow
3.4	Archive synchronization	Transfers data between instances of classes, verifies readability, acknowledges transactions	Archive manager
3.5	DB synchronization	Synchronizes Ops inventory and Operational data stores for retrieval redundancy	RDBMS
3.6	Email notification service	SMTP, MIE, S/MIME-based mail system for notifications	Workflow manager User profile DB
3.7	External system API service	Secure interface support for system-to-system interactions (e.g., order management)	
3.8	FTP servers	In different security regions for various process support	

4.2.2.4 Services Layer

The back-end, batch-oriented processes of CLASS are implemented in the services layer, the internal component-based data processing realm. Data received for acquisition is ingested with QA, metadata extraction, and inventory recording. Workflow-triggered components provide any post-receipt processing and store the data in the archive and the operational data store. Other components are responsible for retrieving data in response to order items, packaging the data for delivery, and invoking interface-layer functionality to perform the delivery.

The archive synchronization processes may be triggered by completion of local archiving or by a scheduling system.

The Oracle Financials order management system is part of this services layer, as is the operational data store.

ID	Component	Description	Interfaces
4.1	Ingest manager	Creates detailed process job control schedule for specific data-stream elements	Inbound data manager Workflow manager
4.2	ObsData QA	Verifies that a received observation dataset is intact (no transmission errors) and meets data-stream-specific quality standards	Job process queues Data quality profile
4.3	Metadata extract	Constructs metadata records for dataset based on data stream, file designation, embedded information, related information	Metadata store
4.4	Inventory record	Creates inventory record (access metadata) for dataset and all ancillary data	Ops inventory
4.5	Browse data generator	Creates metadata (e.g., browse images) from ObsData based on data-stream-specific process descriptions	Operational inventory Data-stream profile Data-stream components
4.6	Data-stream components	Specialized components (provided by data-stream owner) for specific processing	Various, including services interface
4.7	ObsData DB loader	Reformats ObsData for RDBMS storage based on data-stream-specific formats, writes DB	Content DBMS Services interface
4.8	Online data manager	Loads ObsData into accessible storage (cache) with appropriate inventory management	ObsData operational datastore
4.9	Archive manager	Writes any permanent records to archive	Archive Ops inventory
4.10	Order manager (OMS)	Order and order-item management system, creates, tracks, records, completes orders, maintains standing orders	Web interface Order API User directory
4.11	User directory	Provides LDAP-based user record storage, supports authentication and role-based security	Various
4.12	Product distributor-electronic	Based on order processing, retrieves datasets and packages them for distribution	OMS Cache/Archive Ops inventory

ID	Component	Description	Interfaces
4.13	Product distributor-physical	Based on order processing, generates “pick” orders and delivery information (e.g., mailing label, messages to distribution clerks)	Web interface OMS Ops inventory
4.14	Product customization	Performs product-defined, order-specific processing such as subsetting or data fusion (may involve data-stream, owner-supplied components)	OMS Cache/Archive Ops inventory Services interface
4.15	Operational inventory manager	Maintains access information for all products, archived elements, in-process data	Various
4.16	Archive test, verification, and migration	Process support for archive quality assurance and transition to new media	Archive manager
4.17	Content DBMS	GIS-enabled DB application for short-term storage of data-item-level content	ObsData DB loader Services interfaces

4.2.2.5 Data Layer

The data layer includes the storage management systems that underlie the services layer components.

ID	Component	Description	Interfaces
5.1	Archive system	Management system for robotic tape library	Archive Manager
5.2	NAS or SAN	Network Attached Storage system or Storage Area Network for online data storage	Various
5.3	SNA	Storage Network architecture hierarchical data manager	Various
5.4	Inbound FTP staging area		
5.5	Outbound FTP staging area		

4.2.2.6 Workflow Layer

The business process logic of CLASS is the sequencing and evaluation of processes operating on data. Workflow is the generic term for the process control mechanisms that accomplish the business logic. Different layers in the architecture may employ different workflow and business logic components. Some business logic is embedded in services, while some is captured in scripts and command sequences.

ID	Component	Description	Interfaces
6.1	Workflow engine	Continually processes job queue and maintains job status information, schedules and initiates job sequences on available process servers	Various
6.2	Job queue service	Accepts, formats, stores job request elements (processing requests)	Various
6.3	Workflow status database	Contains active job information (status, parameters, file pointers, etc.) as well as workflow log information	
6.4	Workflow edit service	Development tools for workflow modification and template creation	

4.2.3 Workflow Example

The sample use-case scenario developed in section 3.5, describing an end-to-end data flow for a non-interactive, ingest-archive-distribution case, can be mapped against the architecture elements of the layered architecture model. Figure 20 shows the use case against the layers involved in the scenario, along with the supplier and customer interfaces as separate "layers." The circled numbers in the figure correspond to the step numbers in the scenario. The Presentation and Application layers have been omitted because this is a non-interactive scenario.

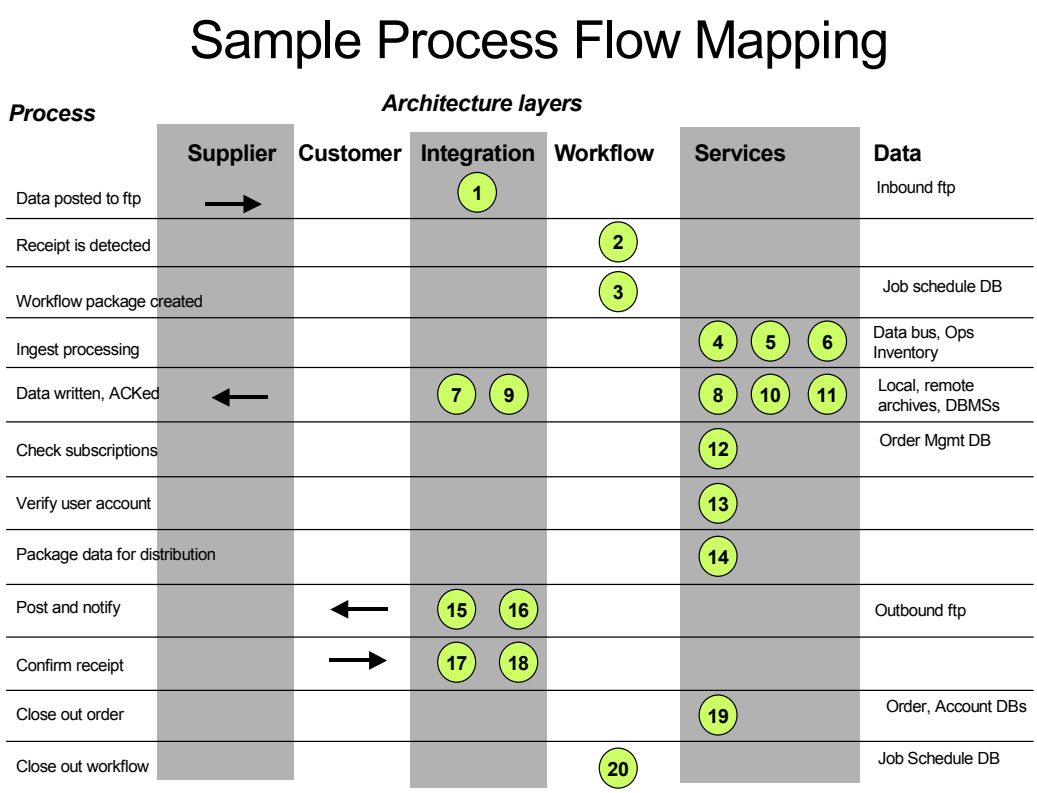


Figure 20. Sample Scenario Mapped to Architecture Model.

The scenario from section 3 is reproduced here and mapped to processes and layers.

#	Scenario Step	Processing Layer/Components
1	Dataset is written to inbound FTP site.	Observation data campaign or program
2	Receipt is detected.	Integration: 3.1 – Inbound FTP area manager
3	Workflow package is initiated.	Workflow: 6.1 – Workflow engine Workflow: 6.2 – Job queue service
4	Ingest processing initiated.	Services: 4.1 – Ingest process
5	QA verification occurs.	Services: 4.2 – ObsData QA
6	Inventory record created.	Services: 4.3 – Inventory record
7	Acknowledgement is sent.	Integration: 3.1 – Inbound FTP area manager (completes process with ack)
8	Local archive is written.	Services: 4.9 – Archive manager

#	Scenario Step	Processing Layer/Components
9	Remote archive is written.	Integration: 3.4 – Archive synchronization
10	Content is formatted for Content DB.	Services: 4.6 – Data-stream components Services: 4.7 – ObsData loader
11	Metadata is written to MDDB.	Services: 4.3 – Metadata extract
12	Standing orders checked; order item created.	Services: 4.10 – Order management system
13	User account is verified.	Services: 4.11 – User directory
14	Data is packaged for distribution.	Services: 4.12 – Product distributor-electronic
15	Data package posted to outbound FTP site.	Integration: 3.3 – Outbound FTP area manager
16	Customer is notified by email.	Integration: 3.6 – email notification service
17	User retrieves package.	Customer
18	Retrieval is detected.	Integration: 3.3 – Outbound FTP area
19	Order item is closed out.	Services: 4.10 – Order management system
20	Workflow package is closed.	Workflow: 6.1 – Workflow engine

4.2.4 Infrastructure

The physical configuration for an instance of CLASS accommodates both the layered software architecture and the mandates of network and system security. The actual server and firewall network depend on security analysis and data center policies. The architecture shown in Figure 21 is a possible secure configuration, placing presentation and outbound FTP service in the least secure segment and incorporating defense in depth to more secure and inward-directed elements.

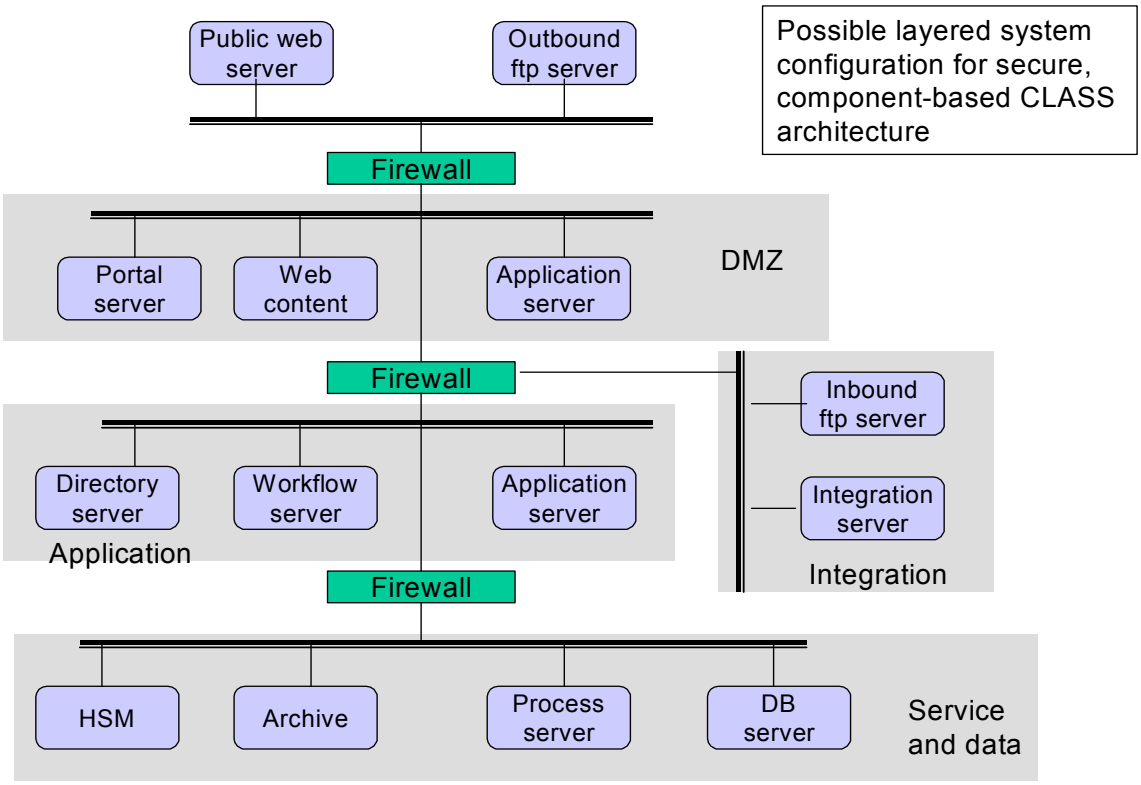


Figure 21. Network and Server Physical Architecture

4.3 Gap Analysis

The current (baseline, "as-is") architecture meets some of the system goals and requirements, but not all, and not in an integrated fashion. This analysis addresses the baseline and target architectures from two frameworks—the expectations and requirements defined in the vision and developed for this architecture and the system characteristics prioritized by the CPMT for CLASS.

It is important to recognize that many of the volume and campaign requirements were not part of the requirements for the SAA (the baseline for CLASS). It is possible that the SAA could have been expanded to support the volumes required with the addition of new data storage capacity. Nevertheless, the baseline does not support these observation data streams, but CLASS does.

The table below discusses the manner in which the two architectures, the baseline and the target architecture, respond to the elements of the CLASS vision (section 2.1). The assumption made for the target architecture is that the major components and functions have been successfully implemented, and where appropriate, a requirement is listed as N/A (not applicable) for the Baseline.

Element	Baseline	Target
CLASS provides access to all necessary NOAA data through a single portal.	NOAA server provides access to many NOAA repositories, but the supporting metadata is difficult to maintain and months out of date.	The discovery portal component, with mechanisms for both automated and manual metadata refreshment, provides access to more and better information, and is more tightly integrated with NESDIS data programs.
CLASS archives, manages, and distributes large-array datasets from current and future observation campaigns.	No. SAA supports a limited set of campaigns but not the CLASS set.	Yes
CLASS provides access and delivery via the Internet.	SAA and the NOAA online web sites separately provide access and delivery for the supported data programs.	CLASS provides an integrated access and delivery mechanism, as well as alternate internet-based integration (e.g., web services, API support).
CLASS supports data discovery and retrieval that is user-needs-centric rather than campaign-centric.	NOAA Server, NVDS, and SAA all provide data discovery, but only at the sophisticated-user level. No straightforward mechanism exists for adding improved discovery and access; all such changes are added piecemeal to the core systems or only at the individual repositories.	CLASS incorporates many of the discovery mechanisms developed for specific programs through support for CLASS-hosted search modules (e.g., templates in the GIS-enabled Content DB) and discovery pass-through to dedicated systems.
CLASS supports novice and inexperienced users as well as intermediate and advanced users from a growing user base.	Limited support on NOAA Server; the SAA interface is geared towards more advanced users; some repository systems provide support (e.g., through SABR).	The CLASS discovery portal provides high-level support for such users, while data program-specific web systems support user-centric, content-specific searching.
CLASS provides secure data storage with multiple storage locations, backup, and disaster recovery capabilities.	Multiple archives in various locations, using off-site storage for disaster recovery, do not accommodate new data campaigns.	Yes, with a common NARA-compliant archive design and standard system management capabilities.
CLASS provides redundant, geographically distributed data processing support for ingest, management, discovery, and distribution.	No	Yes, with systems in Suitland and Asheville.
CLASS reduces the need for new observation campaigns to construct unique data processing systems by providing common information services.	No common services support	Yes, using web services and standard architectures. Implementation requires changes in current systems.

Element	Baseline	Target
CLASS supports new campaigns as they become operational and incorporates existing data systems (directly or with managed access) as priorities and resources dictate.	N/A for new satellite campaigns; no mechanism for incorporating existing data systems.	Yes, with specific policies and procedures for incorporating independent repositories and consolidating common functions.
CLASS supports vastly increased data volumes (compared with current SAA operations) with capacities in the petabyte range.	Yes	Yes; see hardware study for details.
CLASS evolves over time to achieve its goals.	Yes	Yes; see roadmap.
CLASS provides secure storage for massive amounts of data (anticipated to approach 15 petabytes by 2015, growing from slightly more than 30 TB (terabytes) in 2002).	Yes	Yes; see hardware study for details.
CLASS provides reliable 24/7 data receipt, evaluation, and acknowledgment for continuing data streams (as specified in negotiated interface control documents with suppliers).	Yes. The SAA provides this level of reliability.	Yes
CLASS supports the GOES, POES including (DMSP), NEXRAD, EOS, NPP, NPOESS, and Metop campaigns.	Yes	Yes
CLASS supports archive and access to in situ observation data.	No	Yes
CLASS supports archive and access to derived products.	Yes (limited)	Yes, with GIS-capable RDBMS support as well as browse imagery and dynamic product generation.
CLASS distributes data on request via appropriate mechanisms.	Yes (limited; essentially FTP and email)	Yes, with direct web query retrieval as well as FTP and API channels.
CLASS facilitates discovery and access to NOAA environmental data.	Limited capability	Yes, with discovery portal, improved metadata collection, and maintenance.

Element	Baseline	Target
CLASS supports an increasingly diverse set of customers as they explore available data and request selected data products.	Limited capability	Yes
Support Section 508	Limited capability	Yes

The table below compares the two architectures in terms of the system characteristics that must be met in the system design. This evaluation is focused only on SAA, as the core baseline of CLASS.

Characteristic	Baseline	Target
Reliability	The baseline system provides dependable services to its user community.	Similar infrastructure; target adds redundancy and more robust data storage.
Availability	The baseline system operational capabilities exceed its service level agreements. The system automatically detects and recovers for errors.	Similar infrastructure; target adds newer, dedicated data storage, redundant access systems.
Usability	The baseline system is designed for advanced users. Novice users may encounter some difficulty learning how to use the system.	Target is designed for a broader range of users.
Extensibility/Extendibility	The baseline system is designed for extensibility. It supports changes in the specification.	Target architecture is highly extensible and standards-based.
Interoperability	The baseline system supports limited interoperability.	Target is designed to support multiple integration channels (data transfer, middleware, web services, web access, API support)
Scalability	Baseline system is designed for easy addition of additional processors; the baseline system is tied to a specific physical architecture.	Target is more modular and flexible, provides ability to add processors and storage.
Modularity	The baseline system is highly modular.	Target architecture is highly modular and standards-based.
Modifiability	Designed for development as an evolutionary system, the baseline system facilitates the incorporation of changes.	Not a baseline design criterion; target architecture is highly extensible and standards-based.

Characteristic	Baseline	Target
Reusability	The baseline system follows Object-Oriented practices that promote reusability.	Standards-based target is more likely to provide reusable elements.
Integrity	The baseline system has limited security checks.	Highly secure, the target system provides protection against unauthorized access and modification.
Manageability	The baseline system requires experienced operators for supporting the management functions of the system.	The target system provides tools, studies, and custom software to simplify the system management functions.
Testability	The baseline system provides facilities for the development of automated test tools.	The target system incorporates test cases at every stage of the development cycle.
Reparability	Modular and modifiable, the system facilitates the repair of defects.	The target system facilitates the repair of defects.
Understandability	Well documented at the subsystem and module level, the baseline system lacks some high-level system description.	Standards-based target builds on current documentation but will include updated information
Flexibility	The baseline system can easily incorporate some new technologies as they become available (e.g. XML). Other new technologies are difficult to incorporate (e.g. GIS databases).	Target architecture is highly extensible and standards-based.
Efficiency	The baseline system is designed for efficiency.	The target system supports system components that perform its designated functions with minimum consumption of resources.
Recoverability	The baseline system provides error detection and recovery functions for all its components.	Target adds redundant archive.
Security	Lacking some security checks, the baseline system is not designed to handle sensitive information.	Target architecture incorporates central user directory and support for role-based access controls as well as more complex firewall support.
Portability	The baseline system can easily be ported to new AIX hardware platforms or new Informix databases. Porting to other Databases and Operating Systems is a difficult task.	Portability between Unix systems is a design criterion for the target system.

Characteristic	Baseline	Target
Deployability	The baseline system is hard to deploy.	The target architecture does not consider Deployability a high-design criterion.
Survivability	The baseline system continues providing essential functions even if some part of the system is down.	Target adds redundant system configuration.

5. DESIGN CONSIDERATIONS

The design for Release 1 of CLASS is documented in SAA/CLASS documents (please refer to "Satellite Active Archive System and Software Description," library.saa.noaa.gov, 2002). As the requirements are specified in more detail for subsequent releases, the design architecture will evolve towards the architecture presented in this document. The topics in this section are presented as a framework (with some annotation) for the evolving detailed system architecture.

5.1 Data Architecture

The elements of data architecture (DA) include:

a. Hierarchical data classification and description
b. Entity-relationship analysis and modeling
c. Data distribution model
d. Data dictionary (element definitions)
e. Standards, practices, and conventions (e.g., naming conventions)
f. Business data rules

The DA relates to the other model elements (process, location, application, technology) to provide a complete description of how data is managed in the CLASS system.

5.1.1 Data Hierarchy

The hierarchical classification presented below is just one way to decompose the data that CLASS uses and manages. The goal in this study is to identify the major categories of data. During system design, analysis and data modeling serve to refine or redefine the classification.

Primary	Secondary	Tertiary
Science data		
	Raw	
	Processed	
		Permanent products
		Dynamic products
	Meta	
		Summary
		Browse
		Catalog input
		Search input
	Applications	

Primary	Secondary	Tertiary
Customer data		
	Demographic	
	Preferences	
	Record of use	
	Subscriptions	
Operations		
	SD Transactions	
		Data ingest
		Data archive
		Catalog update
		Mirroring
	Order fulfillment	
		Orders
		Order status
		Orders records
		Distribution logs
	QA/QM records	
	Data access and use records	
	Maintenance logs	
Catalog		
	Fileset index	
	Search indexes	
	Summary data	
Security		
	Roles	
	Privileges	
Marketing		
	Data availability studies	
	Data assessments	
	Campaign news	
	CLASS news	

The hierarchy does not attempt to capture the varieties of observation data that are managed through CLASS. In order to provide a stable classification, the schema tries to identify elements that persist. Each new collection of science observation data added to CLASS alters the detailed data classification, so that classification is part of the design rather than the architecture.

5.1.2 Business Process Rules

The relationships among data elements are described by *business process rules*. These rules define the decision-making for each class of data. The business rules also establish policies for data handling (e.g., security policy, retention policies, cost recovery). Business rules (aka *business logic*) are typically embedded in the software of a system in one form or another. Older, legacy systems incorporated business logic directly in the code for data handling. This method is efficient and effective, but expensive to maintain. Somewhat newer systems use table-driven processing controls that permit rules to be updated without recompiling or rewriting application code. These systems often include UIs to permit operators to edit the rules base directly. The most modern approach is to employ a business rules engine (BRE) to move operating logic into COTS (or GOTS) software. The BRE acts as a server with an API with which application code interacts and provides its own rules database and user interface. The BRE may be integrated with a workflow management engine to provide centralized process management.

Business data rules for CLASS can be grouped into the following categories:

a. Security policies: access, sensitivity, security-based retention/destruction, etc.

b. Privacy policies: primarily related to customer account data management.

c. Hierarchical storage migration and accessibility: rules for placing data in online, near-line, or offline storage, and for initiating change of status (e.g., “after four weeks migrate from online to near-line”).

d. Data distribution: specifies the primary and secondary data locations for archived data and locations for directory and other metadata.

e. Backup and retention (e.g., aging, refreshment) policies.

f. Metadata generation: what metadata is required for each archived data type?

g. Product generation: what products are generated from raw data, and by what applications?

h. Quality control process standards and workflow: criteria, evaluation process, steps to follow after pass or failure.

i. Metrics data definitions and studying: what to measure, how, and when to study on the collected data (i.e., scheduled studying).

Other categories may be developed as the analysis continues. The categories serve to organize the rules applying to data; other rules define other operational aspects of CLASS such as user interface interaction standards.

5.2 Technology

Many aspects of this architecture are amenable to implementation with commercial components. A major focus of technology investigation during system design is to evaluate best practices in COTS systems to determine the most effective approach. In some cases, there will be no reliable

commercial products to accomplish CLASS tasks. In other cases, the COTS solutions may simply be too expensive for the value provided. Only specific evaluations determine the best solution. Some of the investigations into commercial product solutions should be:

a. Workflow capabilities
b. Web portal systems
c. Web content management and publishing
d. Middleware (application integration) systems
e. Metadata management systems
f. Search engines (especially in a metadata context)
g. GIS-enabled database systems
h. DBMS synchronization tools
i. User directory services
j. Security models and systems support
k. Data compression systems
l. Multimedia data support (e.g., streaming video)

Although not exhaustive, the list does address the major functional elements for which COTS packages are known to exist.

5.3 Applications

One of the keys to effective open-system support is the ability to include new applications without major design transformation. The e3 architecture framework is very flexible in terms of adding new applications to a process. Interfaces can be defined in a standard form or wrapped in (for example) a J2EE component. Adapters can translate data structures to make common data elements visible. Application interactions can be defined in the workflow process. The amount of custom development needed to support application integration is small compared to the standard architecture. CLASS developers take advantage of this flexibility to include research-developed and commercial components without disruption to the system.

6. TRANSITION PLANNING

Developed in an evolutionary manner, CLASS focuses at first on capturing and archiving major observation campaigns for which no other repository exists and adding integration with other information systems over time and as resources are available. The transition from the baseline CLASS to the target is planned to occur in stages. Timelines are established as part of regular NESDIS budget and planning activities.

A detailed transition plan will be presented in a separate document. The following is a tentative plan of a four-stage transition to the target architecture as defined in this document.

The stages of transition focus on: preparing for growth, expanded data access, a multipurpose portal capability, and a web- and middleware-based services framework.

Stage 1: Preparing for Growth. Stage 1 extends existing system capabilities especially in the area of HMS/Robotic and processor technology, E-commerce capabilities, and archive mirroring.

Stage 1 builds on existing systems to provide the storage capacity and network infrastructure to support the primary large-array-data archive capability. Advanced hierarchical storage management provides more than 5 terabytes of online storage, while new robotic tape storage provides more than a petabyte and is expandable to several petabytes. The SAA job control system serves as the Stage 1 workflow capability. A dual archive capability will be implemented with tape libraries in Suitland, MD and Asheville, NC.

The existing NCDC-based online order management system will be replaced, in part, with the Oracle-based COTS financial system in the COAST project. This is the starting point for e-commerce services that CLASS provides to external NESDIS repositories. Stage 1 implements the OMS/Archive, Access, and Distribution interface.

Stage 2: Expanded Data Access. Stage 2 adds a common user directory, web content management, Content DB (+GIS), enhanced directory-level search, and API access to discovery, ordering, and retrieval.

Stage 2 upgrades the portal capabilities presently delivered with the NOAA Server. A user directory service will be added to provide a single-user profile and authentication capability. A new portal (possibly a COTS application) provides customization and a central web service capability. A web content management capability (possibly COTS) makes it simpler to publicize new data products and NOAA/NESDIS capabilities. The portal provides improved data discovery through enhanced metadata collection and more integrated data-stream information.

Online data support will be improved with the integration of a Content RDBMS supporting GIS functionality, probably with the Informix Geospatial DataBlade capability. User access and searching include data-item-level support. Data managers will be able to define DBMS-resident data logic to perform operations on the online data. Eventually users will be able to use API interfaces to directly access current information.

Stage 3: Multi-Purpose Portal. Stage 3 integrates external repositories, expands content-based search, supports customization, and supports plug-in research-based discovery tools.

Mechanisms will be developed to distribute computing and data processing from CLASS to research data systems, providing a nearly seamless integration of information access. Research-developed processing modules, built to CLASS-specified interface standards, will augment the generic processing capabilities. Content-based search and data fusion capabilities will be expanded in the portal so that archived information becomes more and more useful. Customer profile information will include personalization information to maximize ease of access to desired data. Notification capabilities will be used to alert customers of pre-requested types of data (e.g., customized standing orders using data fusion capabilities).

Stage 4: Services Framework. Stage 4 includes improved workflow/process management, system-system integration, portal-based search, expanded metadata (replace NOAA Server) description, discovery, and retrieval capabilities, and integrated order management.

Stage 4 sees the transition to (for example) a J2EE-based services model for application support, while high-volume transaction processing remains in the workflow-managed, batch-processing domain. Service definition and publication will be developed and coordinated with data center developers to minimize the redevelopment of common functions. The portal will be expanded to provide more complex data discovery as NESDIS metadata initiatives advance. The NOAA server functionality will be completely replaced.

The interfaces with the order management system will be expanded to provide integrated e-commerce support. Access to the Oracle applications will be provided via web service interfaces available to external repositories.

Based on available resources and the requirements of other observation data campaigns, the CLASS System Engineering Team defines the detailed implementation strategy for these four stages.

Appendices

A-1 Vision statements

NESDIS Information Technology Architecture (ITA) (draft)

5.2.1 Comprehensive Large Array-data Stewardship System (CLASS):

CLASS will be a unified, seamless data and information delivery system designed to ensure quality service to users, which include government (federal, state, local), public and private sectors, private citizens, universities and cooperative institutes, and academia. It is the vehicle to be used for the NOAA cross cutting effort to archive high volumes (petabytes) of environmental data and information and to provide access to that data that is critical to the United States Global Change Research Program, the scientific community, and other government agencies as well as NOAA Offices.

5.2.1.1 Goals

The infrastructure improvements required to achieve the CLASS goals are:

1) a high-bandwidth telecommunications infrastructure to convey data from database computers to backbone Internet telecommunications networks;

2) highly-capable enterprise computers to service expanding user demand, including library services and e-commerce; and

3) storage system capable of providing rapid data retrieval as well as high capacity storage.

4) an efficient e-commerce access system.

The goal of the CLASS is to make all archived data available on-line or near on-line through the Internet within practical limits. The CLASS concept is a suite of information services founded on the linking of NOAA's archives with databases at NOAA's laboratories, Regional Climate Centers, State Climate Centers, and other information systems of NOAA's Line Offices. The NOAA/NESDIS three Data Centers are the Nation's stewards of the largest and most comprehensive collection of environmental data and information in the world. The Data Centers holdings represent the chronology of the Nation's environmental history, as well as records from many other regions of the world. A large portion (over 750 terabytes) of the Nation's archive of environmental data is stored and maintained by the three data centers and the Satellite Active Archive (SAA), which are located in different areas of the country. Ultimately, this concept will also archive and provide access to smaller and unique environmental datasets from diverse providers across the nation. CLASS will provide straightforward, easy access to NOAA-managed environmental data, information, and products to a widely diverse, worldwide clientele.

5.2.1.2 Challenges:

1) The storage and access of large volumes of new data that increases in one year by the equivalent of all the data NOAA has had to handle over the past 100 years - The storage and access of large volumes of historical data which are currently difficult to access

2) Making environmental data easily accessible to large numbers of new and inexperienced users who are using the Internet to access data.

3) The sources of the vast volumes of new data are:
   - Existing and planned from NOAA, NASA, and DoD observing systems
   - Earth Observing System (EOS)

4) Other sources of vast volumes of data are:
   - Change Research Program, the scientific community, and other government agencies as well as NOAA Offices.

5.2.1.3 Requirements:

The goal of the CLASS is to make all data available on-line or near on-line through the Internet within practical limits. The CLASS concept is a suite of information services founded on the linking of NOAA's archives with databases at NOAA's laboratories, Regional Climate Centers, State Climate Centers, and other information systems of NOAA's Line Offices.

In preparation for these data, NOAA must:

1) Improve ability to ingest large amounts of data (to over a TB per day by 2004)
2) Increase data-handling capacity and capabilities of its Data Centers
3) Expand its current NOAA/NASA short-term archive project
4) Achieve rapid expansion in storage capacity at the Data Centers
5) Achieve automating the means of data ingest, quality control, and access
6) Improve the e-commerce capabilities

5.2.1.4 CLASS Impact on NESDIS IT Domains

5.2.1.4.1 Network Domain

Improve communications facilities. The NOAA National Data Centers (NNDC) will increase their communications pipeline to an OC-12c (622Mbs) high capacity line or better. The OC-12c (and later OC48 (248 Gbs)) will be able to handle high resolution, large array datasets such as weather radar data (NEXRAD), weather satellite data (POES, GOES and EOS), and environmental data from DMSP. In addition, the NNDC will upgrade their current network, routers, communication servers, and workstations. This will allow smoother data ingest, and also ease access to the NNDC data by researchers, educators, and commercial users.

5.2.1.4.2 Archive Management Domain

The NNDC has approximately 2 terabytes of data on-line, 12 terabytes near on-line, and 725 terabytes off-line. To improve access to the data in the archive (off-line data), the NNDC will expand their mass storage robotics systems to one and a half petabytes (10 3 terabytes or 10 6 gigabytes). This and associated information technology are required to access the data in a more timely manner. The NNDC will also increase their computer processing power. This will enhance quality control and reprocessing of high volumes of data. The data can then be put in historical perspective for use by researchers and decision-makers. The NNDC will procure data compression technology to compress and decompress these important data

5.2.1.4.3 Product Generation Domain

To monitor near real-time climatic, oceanic, seismological, and environmental events, the NNDC will electronically ingest and archive all available high resolution NOAA observational and satellite datasets. High volume, high-resolution scientific data are crucial to quantifying the risk of weather- and other

hazards-related disasters at community levels. Implement tools to allow users to extract data from on-line databases and to generate customized products “on-the-fly”.

5.2.1.4.3.1 Combine datasets and produce visualizations

The NNDC will use advanced technology visualization software and expertise in integrating (fusing) multiple datasets into useful information. Data fusion and visualization represent new areas within the scientific data community that the NNDC have not been able to fully utilize for lack of resources. Instantaneous access to data and the visualization tools lead to improved environmental issues by businesses, researchers, and decision and policy makers.

5.2.1.4.4 Research and Development Domain

The NNDC will enable broader access to past and present data records necessary for developing new algorithms for characterizing data and understanding long term impacts tothe environment using historical data bases.

5.2.1.4.5 Services Domain

The greatest use of the NNDC data (oceanographic, atmospheric, geophysical) is by the government and private sectors, nationally and internationally, to control and manage hazards risk. These users depend upon the NNDC ability to assess the probabilities of extreme events. With increased data resolution, improved computing performance, additional geographical information systems, and state-of-the-art data visualizations, the NNDC will apply new information technology to prepare new user-based products. The NNDC leadership possesses unique capabilities to assure that U.S. interests are maximized while international laws and agreements are honored.

5.2.1.4.5.1 Combine datasets and produce visualizations.

The NNDC will use advanced technology visualization software and expertise in integrating (fusing) multiple datasets into useful information. Data fusion and visualization represent new areas within the scientific data community that the NNDC have not been able to fully utilize for lack of resources. Instantaneous access to data and the visualization tools lead to improved environmental stewardship by businesses, researchers, and decision and policy makers.

5.2.1.4.5.2 Improve analysis capabilities.

The NNDC will procure additional Geographic Information System (GIS) software and expertise to improve analytic capabilities. NOAA data can be used within a GIS to identify areas of severe weather, geophysical activity (such as volcanoes and earthquakes), fire, extreme biological events, and coastal hazards, including tsunamis. GIS enables more rapid risk assessments.

CIO - Vision statement (NESDIS CIO - April 24, 2002)

The goals of CLASS are:

1. Give any potential customer access to all NOAA (and possibly non-NOAA) data through a single portal
2. Eliminate the need to keep creating "stovepipe" systems for each new type of data, but in as much as possible use already polished portions/modules of existing legacy systems
3. Describe a cost-effective architecture that can primarily handle large array data sets but also be capable of handling smaller data sets as well

To meet these goals the CLASS system will unfold over a period of time. A phased approach will be used to keep cost, per year, to a minimum. The first phase of CLASS will be to design and build an access portal capability based upon the existing Satellite Active Archive (SAA). During the first phase of development this front end system will be attached to the back end systems at the SAA and installed in Asheville, NC and attached to their existing legacy storage systems. This will provide access to both NOAA POES and GOES data from two separate sites. As part of this phase a robust distributed development environment will be established and a robust configuration management discipline will be implemented. The next phase of the CLASS development embraces the DoD DMSP data sets and the functionality of the NGDC developed legacy support systems. The ensuing phases will be issued as enhanced versions to the CLASS system adding new functionality and additional data sets. These phases are important phases as it allows the customers of CLASS to obtain the data they want, even though there may still be separate "stovepipe" back end systems providing the data, through the single CLASS portal. The following phases will go on without the customer knowing about them, in other words, no loss of service, only an increase where possible. The CLASS user interface will be able to allow a novice to be lead through the various types of data they may be interested in. It will also be able to aid the more advanced (intermediate) customer to cut through steps to get to the type of data they are looking for, and it will have even more short cuts for the advance customer that knows what they want. This interface needs to also allow for basic analyzation of data to aid the customer in choosing what to order (maybe).

The CLASS portal is the user's vision of CLASS but the CLASS is much more. CLASS will contain software that will properly interpret a query from the portal and be able to quickly discover the data type(s) that the customer is interested. This can be done through a networked relational database management system (RDBMS). The database(s) will identify all CLASS sites which contain the data (mirroring) in case of any network outages encountered when querying the closest site.

CLASS starts out with the portal into existing legacy systems and then phases in each of the legacy systems into the primary CLASS facilities, as well as providing the tools and configuration management discipline for future systems to be built into the CLASS application. The goal for CLASS is to build one, object oriented, application software system, which will be able to ingest, QA, archive and distribute all data types, adding one or more at a time. This CLASS application will be maintained by one team, taking change requests from at least one backup CLASS site and providing updates in the form of version releases similar to the way vendors today provide COTS tools and version updates to the same, via CD or internet downloads.

Over time there will be a cost savings as maintenance of all of the “stovepipe” systems will dissolve into the one CLASS maintenance. The fact that CLASS can access data from anywhere, through the portal, will allow NOAA management to physically locate the primary and backup CLASS sites where they will be the most economic.

A-2 National Archive And Records Administration (NARA) guidelines

The information below is extracted from NARA regulations part 1234, and can be found at http://www.archives.gov/about_us/regulations/part_1234.html

§ 1234.30 Selection and maintenance of electronic records storage media.

(a) Agencies shall select appropriate media and systems for storing agency records throughout their life, which meet the following requirements:

(1) Permit easy retrieval in a timely fashion;

(2) Facilitate distinction between record and non-record material;

(3) Retain the records in a usable format until their authorized disposition date; and

(4) If the media contains permanent records and does not meet the requirements for transferring permanent records to NARA as outlined in 1228.270 of this chapter, permit the migration of the permanent records at the time of transfer to a medium which does meet the requirements.

(b) The following factors shall be considered before selecting a storage medium or converting from one medium to another:

(1) The authorized life of the records, as determined during the scheduling process;

(2) The maintenance necessary to retain the records;

(3) The cost of storing and retrieving the records;

(4) The records density;

(5) The access time to retrieve stored records;

(6) The portability of the medium (that is, selecting a medium that will run on equipment offered by multiple manufacturers) and the ability to transfer the information from one medium to another (such as from optical disk to magnetic tape); and

(7) Whether the medium meets current applicable Federal Information Processing Standards.

(c) Agencies should avoid the use of floppy disks for the exclusive long-term storage of permanent or unscheduled electronic records.

(d) Agencies shall ensure that all authorized users can identify and retrieve information stored on diskettes, removable disks, or tapes by establishing or adopting procedures for external labeling.

(e) Agencies shall ensure that information is not lost because of changing technology or deterioration by converting storage media to provide compatibility with the agency's

current hardware and software. Before conversion to a different medium, agencies must determine that the authorized disposition of the electronic records can be implemented after conversion.

(f) Agencies shall back up electronic records on a regular basis to safeguard against the loss of information due to equipment malfunctions or human error. Duplicate copies of permanent or unscheduled records shall be maintained in storage areas separate from the location of the records that have been copied.

(g) Maintenance of magnetic computer tape. (1) Agencies shall test magnetic computer tapes no more than 6 months prior to using them to store electronic records that are unscheduled or scheduled for permanent retention. This test should verify that the tape is free of permanent errors and in compliance with National Institute of Standards and Technology or industry standards.

(2) Agencies shall maintain the storage and test areas for computer magnetic tapes containing permanent and unscheduled records at the following temperatures and relative humidity's:

Constant temperature -- 62 to 68oF.
Constant relative humidity -- 35% to 45%

(3) Agencies shall annually read a statistical sample of all reels of magnetic computer tape containing permanent and unscheduled records to identify any loss of data and to discover and correct the causes of data loss. In tape libraries with 1800 or fewer reels, a 20% sample or a sample size of 50 reels, whichever is larger, should be read. In tape libraries with more than 1800 reels, a sample of 384 reels should be read. Tapes with 10 or more errors should be replaced and, when possible, lost data shall be restored. All other tapes which might have been affected by the same cause (i.e., poor quality tape, high usage, poor environment, improper handling) shall be read and corrected as appropriate.

(4) Agencies shall copy permanent or unscheduled data on magnetic tapes before the tapes are 10 years old onto tested and verified new tapes.

(5) External labels (or the equivalent automated tape management system) for magnetic tapes used to store permanent or unscheduled electronic records shall provide unique identification for each reel, including the name of the organizational unit responsible for the data, system title, and security classification, if applicable. Additionally, the following information shall be maintained for (but not necessarily attached to) each reel used to store permanent or unscheduled electronic records: file title(s); dates of creation; dates of coverage; the recording density; type of internal labels; volume serial number, if applicable; number of tracks; character code/software dependency; information about block size; and reel sequence number, if the file is part of a multi-reel set. For numeric data files, include record format and logical record length, if applicable; dataset name(s) and sequence, if applicable; and number of records for each dataset.

(6) Agencies shall prohibit smoking and eating in magnetic computer tape storage libraries and test or evaluation areas that contain permanent or unscheduled records.

(h) *Maintenance of direct access storage media.* (1) Agencies shall issue written procedures for the care and handling of direct access storage media which draw upon the recommendations of the manufacturers.

(2) External labels for diskettes or removable disks used when processing or temporarily storing permanent or unscheduled records shall include the following information: name of the organizational unit responsible for the records, descriptive title of the contents, dates of creation, security classification, if applicable, and identification of the software and hardware used.